Project Report

LMA Major Project

Course Name : Language Models and Agents
Course Code : CL3.410
Semester : Monsoon'25
Instructor Name : Prof. Vasudev Varma & Vandan Majudia

Pratyusha Mitra (2024701033)
(pratyusha.mitra@research.iiit.ac.in)

Kushagra Gupta(2025909001)
(kushagra.g@students.iiit.ac.in)

International Institute of Information Technology
Hyderabad, India

October 25, 2025

Contents

1 Data Collection and Preparation

1.1 Phase 1: Document Collection and Organization

The document collection phase involved sourcing authoritative textual material relevant to the domain of **Food Safety and Nutrition**. A representative NCERT textbook, `12_homescience_eng_sm_2024.pdf`, was used as the initial data source to establish and test the complete question–answering pipeline.

All source documents were organized under a root directory within Google Drive. Each file was automatically detected and processed using the `PyPDFLoader` module from the `langchain-community` library. For every extracted document, a metadata schema was attached to facilitate structured indexing and retrieval. The metadata included attributes such as:

- **Subject:** Food Safety and Nutrition
- **Source:** NCERT Textbook
- **Timestamp:** ISO-8601 formatted ingestion time

This ensured that each chunk or retrieved passage could be traced back to its source context during later stages of retrieval and analysis.

The setup environment was configured in Google Colab, with the following dependencies installed:

```
!pip install -q langchain langchain-community langchain-openai
               sentence-transformers faiss-cpu pypdf
```

Google Drive was mounted to provide persistent storage and facilitate index reuse.

The project titled `food_safety_rag` was developed as a modular retrieval-augmented generation (RAG) system for the domain of **Food Safety and Nutrition**. Two primary data sources were used for corpus construction:

1. **NCERT Home Science Textbook (Class XII, 2024 edition):** extracted from `12_homescience_eng_sm_2024.pdf` using the `PyPDFLoader` utility.
2. **Hugging Face Dataset:** the publicly available corpus `yasserrmd/food-safety` was imported using the `datasets` library to supplement textbook material with contemporary food-safety-related documents.

All collected material was stored in a structured repository (`food_safety_rag/`) consisting of separate subdirectories for data, scripts, and configuration files. Each document was annotated with essential metadata fields, namely *subject*, *source*, and *timestamp*, to support traceability and provenance management. The repository follows the following layout:

```
food_safety_rag/
  data/
```

```
    food_safety_texts.txt
    combined_corpus.txt
    dataset.py
scripts/
    build_database.py
    retriever.py
    query_rag.py
    test_es.py
requirements.txt
docker-compose.yml
venv/
```

Dependency management was defined in `requirements.txt`, which included core libraries such as `transformers`, `sentence-transformers`, `datasets`, `pypdf`, and `elasticsearch`: contentReference[oaicite:1]index=1. All components were executed within a virtual environment and verified via shell scripts (`setup_rag.sh`, `build.sh`, `run.sh`) to ensure reproducibility.

1.2 Phase 2: Document Collection and Organization

- All source materials were collected under a root-level directory in Google Drive (`/food_safety_corpus/`). This directory contains heterogeneous files including PDF, DOCX, PPTX, TXT, and Markdown formats.
- The ingestion pipeline automatically detects document types and assigns appropriate loaders:
  - `PyPDFLoader` for PDF documents
  - `UnstructuredWordDocumentLoader` for DOCX files
  - `UnstructuredPowerPointLoader` for PPTX presentations
  - `TextLoader` for TXT files
  - `UnstructuredMarkdownLoader` for MD files
- Each document is enriched with metadata attributes to preserve provenance and structure:
  - **subject:** Food Safety and Nutrition
  - **source:** File name or dataset origin
  - **context:** Authoritative or supplementary reference
  - **timestamp:** Ingestion time in ISO format
  - **file_type:** Original document format
- The system integrates two complementary data sources:
  1. The NCERT textbook `12_homescience_eng_sm_2024.pdf`, processed with `PyPDFLoader`.
  2. The Hugging Face dataset `yasserrmd/food-safety`, imported through the `datasets` library to supplement the textbook content with real-world examples.
- A total of 748 pages were processed from the primary corpus, and all content was serialized into a unified document list with consistent metadata.

- The pipeline supports scalable addition of new authoritative or supplementary sources while maintaining consistent structure and traceability.

2 Preprocessing and Chunking

The preprocessing pipeline began by extracting raw text content from the PDF and applying a cleaning function to normalize case, remove extraneous symbols, and collapse redundant whitespace. All text was converted to lowercase and stripped of non-informative characters using regular expressions.

Document segmentation was performed using the `RecursiveCharacterTextSplitter` class from `LangChain`. This strategy allows content-aware chunking with adjustable granularity and overlap, maintaining semantic continuity across segments. The configuration used for this project was:

- `chunk_size = 1000` characters
- `chunk_overlap = 200` characters

This resulted in a collection of self-contained text chunks suitable for embedding and retrieval. The recursive splitter ensures that paragraph and sentence boundaries are respected wherever possible, reducing information fragmentation.

The data preprocessing pipeline was implemented both interactively (in Colab) and through the script `scripts/build_database.py`. It performs the following key operations:

1. **Text Normalization:** lowercasing, whitespace normalization, and removal of non-alphanumeric characters via regular expressions.
2. **Chunking Strategy:**content-aware segmentation using the `RecursiveCharacterTextSplitter` with parameters `chunk_size = 1000` and `chunk_overlap = 200`.
3. **Token-level Cleaning:** deduplication and filtering of empty or non-informative text segments.

The same logic was applied to both locally extracted text and data loaded from the Hugging Face dataset. All processed segments were concatenated into unified corpus files (`food_safety_texts.txt` and `combined_corpus.txt`) inside the `data/` directory. This ensured semantic continuity and minimized loss of contextual information between consecutive segments.

2.1 Preprocessing and Chunking

- Text preprocessing was applied uniformly to both the NCERT textbook and the Hugging Face dataset. Each document underwent normalization to remove noise, convert text to lowercase, and collapse redundant whitespace.
- The following regex-based cleaning operations were implemented:
  - Removal of URLs, HTML tags, and non-alphanumeric symbols.
  - Tokenization and lowercasing for consistent case handling.

  - Filtering of non-informative or very short segments.
  - Deduplication via MD5 hashing of cleaned text.
- The base chunking strategy adopted was **content-aware recursive splitting** using the `RecursiveCharacterTextSplitter`. Parameters were tuned as:
  - `chunk_size = 1000` characters
  - `chunk_overlap = 200` characters

  This ensured semantic continuity across paragraph boundaries.
- To facilitate hierarchical retrieval, **multi-granularity segmentation** was implemented using a token-based approach with the Hugging Face `bert-base-uncased` tokenizer:
  - `2048`, `512`, and `128`-token windows
  - Approximately 15% token overlap per level
- The pipeline was designed as a batch ingestion process (`process_documents_batch`) with progress tracking via `tqdm` and logging to `ingestion.txt`.
- All chunk levels (`content_recursive`, `lvl_2048`, `lvl_512`, `lvl_128`) were stored in a structured dictionary, later serialized for embedding.
- Comprehensive error handling and logging were implemented to capture ingestion anomalies and record per-stage statistics for reproducibility.

3 Embedding and Indexing

3.1 Embedding and Indexing

- A **parent–child hierarchy** was established to maintain document structure. Each source document (parent) was assigned a unique `parent_id`, and all derived text chunks (children) reference this identifier to preserve source linkage.
- Two embedding models were employed:
  - **General model:** `sentence-transformers/all-MiniLM-L6-v2` (384 dimensions) for baseline semantic coverage.
  - **Domain-specific model:** `allenai/specter2_base` (768 dimensions) tuned for scientific and educational corpora in food safety and nutrition.
- Embeddings were computed in optimized batches with GPU auto-detection and caching (`.npy` files) to minimize recomputation. Each embedding was normalized for cosine similarity search.
- Two FAISS vector indices were constructed:
  - **FAISS (General):** Built with MiniLM embeddings for fast, low-dimensional retrieval. [1]
  - **FAISS (Domain):** Built with SPECTER2 embeddings for domain-aware retrieval.

  Both indices were persisted in Google Drive under `/fsn_outputs/faiss_general` and `/fsn_outputs/faiss_domain`.
- Metadata for each chunk, including the parent identifier, granularity level, and file source, was serialized into `meta.pkl` files for structural integrity.
- A **retrieval and reranking module** was implemented:

    - Initial candidate retrieval from FAISS using L2 similarity.
    - **BGE Reranker (BAAI/bge-reranker-base)** applied for semantic relevance scoring.
    - Fallback to FAISS similarity if reranker is unavailable.
- The reranking step assigns contextual scores to query–document pairs, improving precision over base vector similarity.

Bonus A local hybrid retriever (FAISS + BM25) was also developed to emulate Elasticsearch-style mixed relevance ranking without requiring a running ES server.

The following code snippet summarizes the process:

```
model = SentenceTransformer("all-mpnet-base-v2")
embeddings = model.encode(texts, show_progress_bar=True, convert_to_numpy=True)
index = faiss.IndexFlatL2(embeddings.shape[1])
index.add(embeddings)
faiss.write_index(index, "faiss_food_safety.index")

with open("metadata.pkl", "wb") as f:
    pickle.dump((texts, metadatas), f)
```

Both the FAISS index and the metadata pickle file were stored in Google Drive for later reuse:

```
/content/drive/MyDrive/fsn_outputs/faiss_food_safety.index
/content/drive/MyDrive/fsn_outputs/metadata.pkl
```

Verification

To validate successful indexing, a retrieval function was implemented to encode a user query, perform a similarity search against the FAISS index, and return the top-k semantically closest chunks. For example, when queried with *"What are the main causes of food spoilage?"*, the system retrieved coherent and contextually relevant passages discussing factors such as microbial contamination, physical and chemical deterioration, and inappropriate storage conditions.

This confirmed that the embedding and indexing pipeline was functioning as intended and provided semantically meaningful retrievals for downstream question-answering tasks.

Semantic embeddings for each text chunk were generated using the `SentenceTransformer` model `all-mpnet-base-v2`. The embedding and indexing workflow was implemented both in the Colab prototype and in the script `scripts/retriever.py`, following these steps:

1. Compute dense 768-dimensional embeddings for all text chunks.
2. Build a FAISS index using the `IndexFlatL2` structure for efficient nearest-neighbor search.

3. Persist the vector index (`faiss_food_safety.index`) and its corresponding metadata (`metadata.pkl`) for later retrieval.

The following code summarizes the process:

```
model = SentenceTransformer("all-mpnet-base-v2")
embeddings = model.encode(texts, show_progress_bar=True, convert_to_numpy=True)
index = faiss.IndexFlatL2(embeddings.shape[1])
index.add(embeddings)
faiss.write_index(index, "faiss_food_safety.index")

with open("metadata.pkl", "wb") as f:
    pickle.dump((texts, metadatas), f)
```

The index and metadata were subsequently loaded from persistent storage using the `faiss.read_index()` and `pickle.load()` utilities. A retrieval function was defined in `retriever.py` and `query_rag.py` to encode user queries, perform top-k similarity search, and return semantically closest text segments for downstream question answering.

Verification

Initial evaluation queries, such as *"What are the main causes of food spoilage?"*, retrieved contextually accurate results covering microbial, chemical, and environmental causes of spoilage, confirming that the embedding and indexing pipeline was correctly implemented.

4 Core SME Capabilities

5 Agent for Chat, Planning/Reasoning, and Routing

6 LLM models and RAG:Retrieval and Relevance Ranking Mechanisms

7 Tool capabilities

8 System Components Architecture

References

[1] M. Douze, A. Guzhva, C. Deng, J. Johnson, G. Szilvasy, P.-E. Mazaré, M. Lomeli, L. Hosseini, and H. Jégou, "The faiss library," 2024.