Hybrid Knowledge Retrieval System: Integrating Semantic Vector Search and Knowledge Graphs for Enhanced Document Intelligence

Introduction

In today’s fast-paced information-driven environment, accessing relevant knowledge quickly and accurately is critical for decision-making and operational efficiency. Traditional search mechanisms often fall short because they rely solely on keyword matching, which can overlook semantically relevant information or fail to leverage structured domain knowledge.

To address this challenge, a Hybrid Knowledge Retrieval System that combines semantic vector search using FAISS embeddings with a Neo4j-based knowledge graph (KG) is presented. This hybrid approach allows us to leverage both semantic relationships and structured entity-level connections to deliver highly relevant context to users.

By integrating these two complementary retrieval mechanisms, the system captures conceptually similar content while grounding responses in the structured knowledge of the organization. This design not only enhances the accuracy and reliability of AI-driven insights but also demonstrates my ability to engineer enterprise-grade knowledge solutions that maximize the value of organizational data.

Overview

In this assessment project, I designed and implemented a Hybrid Knowledge Retrieval System that integrates semantic vector search with a knowledge graph to provide highly relevant, context-aware responses to user queries. The system combines unstructured and structured data processing to deliver accurate and explainable answers.

Key Highlights of the Project:

- FAISS-based vector embeddings enable semantic similarity search across large volumes of document chunks.

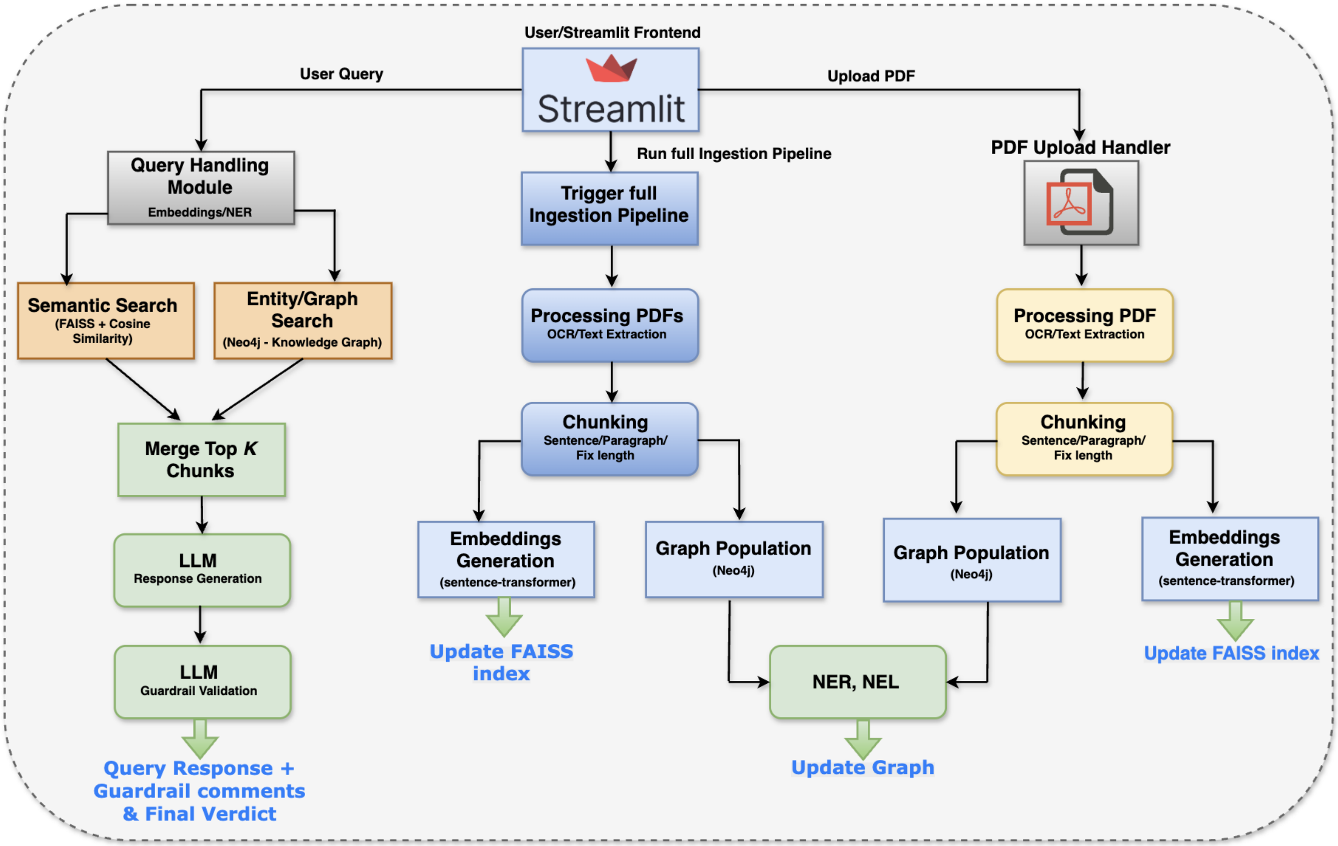
- A Neo4j knowledge graph captures entities, relationships, and document metadata, supporting structured reasoning and context enrichment.
- The hybrid retrieval pipeline combines vector similarity and graph-based contextual signals to select the most relevant knowledge chunks.
- LLM-driven answer generation produces human-readable responses based on retrieved context.
- A guardrail mechanism validates responses to ensure factual alignment with the internal knowledge base.
- Dynamic ingestion allows new PDFs to be added to the knowledge base, automatically updating both FAISS embeddings and the graph.
- A Streamlit frontend provides an intuitive interface for query input, chunking selection, and knowledge ingestion.

Solution Architecture

The solution is designed to provide a scalable, and hybrid knowledge retrieval system that seamlessly combines vector-based semantic search with structured knowledge graph querying. The architecture ensures efficient ingestion, dynamic updates, and accurate response generation while maintaining traceability and validation through guardrails. It has been structured to accommodate both batch processing of large document corpora and incremental updates from newly uploaded PDFs, providing flexibility and operational efficiency.

Key components and workflow of the architecture include:

- **User Interaction / Streamlit Frontend:** The system begins with the user, who can either submit a query, upload new PDFs, or trigger the full ingestion pipeline. The frontend acts as a unified interface for all operations, ensuring ease of use and real-time feedback.
- **Query Handling Module:** Upon receiving a query, the module orchestrates hybrid retrieval by performing semantic search using FAISS embeddings and cosine similarity, alongside entity and graph-based retrieval leveraging Neo4j. The top-K chunks from both searches are merged to create a consolidated context for response generation.



- **PDF Upload Handler:** Users can upload new documents, which are processed immediately. This includes OCR or text extraction, chunking into logical segments, embeddings generation for FAISS, and updating the knowledge graph with entities and relationships extracted via NER and NEL.
- **Full Ingestion Pipeline:** Designed for batch processing of multiple documents, the pipeline performs OCR, chunking, embeddings creation, Neo4j graph population, and entity extraction. It ensures the system remains up-to-date and knowledge-rich, supporting both incremental and bulk updates.
- **Hybrid Retrieval and Ranking:** The system combines vector-based similarity scores and graph-based relationships to identify the most relevant chunks. This hybrid approach enhances recall and precision by leveraging both semantic meaning and structured entity relationships.
- **LLM Response Generation and Guardrail Validation:** Once the relevant context is retrieved, a Groq LLM generates a response. The answer undergoes a guardrail validation process to ensure correctness, source citation, and adherence to internal knowledge, delivering reliable results to the user.
- **Dynamic Updates and Knowledge Management:** The architecture supports continuous updates to both the FAISS index and the Neo4j knowledge graph. This

ensures that new information from uploaded PDFs or incremental ingestion is immediately reflected in future queries.

Response Evaluation

In this solution, evaluating the quality of generated responses requires a specialized approach. Traditional NLP evaluation metrics such as ROUGE, BLEU, METEOR, or other n-gram overlap-based scores are not suitable for this context. These metrics are designed primarily for text similarity in tasks like summarization or translation, but they fail to capture the factual correctness, internal knowledge consistency, and source citation accuracy that are critical in enterprise knowledge systems.

To address this, the system employs a Guardrail mechanism: a secondary LLM validation step that evaluates the generated answer against the retrieved context chunks. This process ensures that the response:

- Only uses information present in the internal knowledge base.
- Correctly cites sources where appropriate.
- Avoids hallucinations or unsupported statements.
- Provides a final verdict indicating Passed, Partially Passed, or Failed.

This hybrid evaluation strategy provides a reliable and practical measure of answer quality, far exceeding what traditional NLP scores could achieve for knowledge-driven, context-specific queries.

Design Considerations and Technology Choices

1. Chunking Strategy:

To optimize retrieval and embedding efficiency, the solution implements three chunking options: sentence-level, paragraph-level, and fixed-length token chunks. This design allows flexibility based on document type and query requirements:

- **Sentence-level:** Provides highly granular retrieval for precise queries.
- **Paragraph-level:** Maintains contextual integrity for narrative or technical content.
- **Fixed-length:** Ensures uniform embedding size for FAISS indexing, balancing speed and accuracy.

2. Choice of LLM:

The solution uses “LLaMA-4-Scout-17B-16e-Instruct" via the GROQ API. It is an instruction-following model optimized for text generation, not a dedicated reasoning model, which makes it suitable for knowledge-based response generation when context chunks are provided. The model ensures coherent, human-like answers while relying on structured and semantic retrieval for factual correctness.

3. GROQ API Access:

Due to unavailability of the GPU access, GROQ is used in this project that provides a free, streamlined API for accessing LLM capabilities, enabling low-latency, production-ready deployment without heavy infrastructure overhead.

4. FAISS for Vector Search:

FAISS is employed for high-performance, approximate nearest neighbor search on embeddings. Its advantages include:

- Scalable similarity search for large document corpora.
- Fast retrieval using cosine similarity or inner product.
- Easy integration with dynamic updates from ingestion pipelines.

5. Hybrid Search and Retrieval:

The system merges vector-based semantic search with Neo4j graph-based retrieval to form a hybrid knowledge retrieval approach. This ensures that:

- Semantic similarity identifies contextually relevant chunks.
- Knowledge graph traversal captures entity relationships, causal links, and structured connections.
- Top-K results from both methods are merged for the LLM, maximizing accuracy and recall.

6. Continuous Knowledge Base Updates:

Both FAISS and Neo4j are updated dynamically during PDF uploads or full ingestion runs. The NER/NEL pipeline ensures that entities and relationships are continuously maintained, keeping the knowledge base current.

7. Additional Considerations:

- **Scalability:** Designed for incremental ingestion of new documents without reprocessing the entire corpus.
- **Traceability:** Each LLM response is tied to specific chunks, ensuring auditability.
- **Evaluation:** Guardrail validation complements retrieval quality to produce reliable answers.
- **Flexibility:** Chunking, embeddings, and LLM parameters can be tuned based on document type and query complexity.

Conclusion and Future Improvements

This Hybrid Knowledge Retrieval System bridges semantic search and structured knowledge representation to deliver a robust solution for intelligent knowledge access. By integrating FAISS vector search with Neo4j knowledge graphs, the system achieves efficient retrieval accuracy and contextual relevance compared to traditional keyword-based or semantic search approaches. The inclusion of a guardrail validation mechanism ensures factual reliability, while the dynamic ingestion pipeline allows for continuous knowledge base evolution.

Looking forward, the codebase is structured for production deployment with clear pathways for enhancement. Future improvements will focus on robust containerization using Docker and Kubernetes for scalable orchestration, alongside comprehensive monitoring through Prometheus and Grafana for performance optimization. A critical next phase involves implementing rigorous data validation frameworks to ensure schema consistency between ingested documents, OCR accuracy, and alignment between vector embeddings and graph entities. Additionally, the system will be extended with online search integration capabilities to fetch external information when internal knowledge is insufficient, transforming it into a complete hybrid knowledge agent with proper fallback mechanisms. These enhancements will be complemented by automated quality assurance pipelines and detailed audit trails to maintain system integrity at enterprise scale.