Agente RAG: Funcionamiento y Pruebas

Grupo 4

Introducción

- **Agente RAG (Retrieval-Augmented Generation)**:
  - Combina recuperación de información y generación de texto.
  - Utiliza modelos de lenguaje como GPT-4 para responder preguntas basadas en documentos.
- **Objetivo**:
  - Responder preguntas utilizando información contenida en archivos PDF.

Flujo de Trabajo del Agente

1. **Carga de documentos PDF**:
   - Los archivos PDF se cargan desde un directorio específico.
   - Se procesan utilizando `PyPDFLoader`.
2. **División del texto**:
   - Los documentos se dividen en fragmentos (`chunks`) utilizando `RecursiveCharacterTextSplitter`.
   - Esto facilita la indexación y recuperación eficiente.
3. **Creación de embeddings**:
   - Se generan representaciones vectoriales (embeddings) de los fragmentos usando `OpenAIEmbeddings`.
4. **Indexación en Chroma**:
   - Los embeddings se almacenan en una base de datos vectorial (`Chroma`).
   - Esto permite realizar búsquedas rápidas basadas en similitud.
5. **Recuperación y generación**:

- El agente utiliza un modelo de lenguaje (GPT-4) para responder preguntas.
- Recupera información relevante de la base de datos y genera respuestas.