MANUAL TÉCNICO DE PROCEDIMIENTO

GUÍA DE IMPLEMENTACIÓN RAG

INTRODUCCIÓN

Crear un sistema RAG implica conectar un modelo de lenguaje con una base de conocimientos externa. A continuación, se detallan los pasos técnicos necesarios para construir una solución robusta y funcional.

PASO 1: CONFIGURACIÓN DEL ENTORNO

Para este proyecto utilizaremos Python como lenguaje base y bibliotecas líderes en el sector:

- LangChain: Orquestador fundamental para conectar las distintas piezas de la IA.
- ChromaDB o FAISS: Bases de datos vectoriales para almacenar la información.
- PyPDF o Unstructured: Para cargar documentos en distintos formatos.

PASO 2: CARGA DE DOCUMENTOS (INGESTION)

El primer paso es leer los archivos que formarán nuestra base de conocimiento. Estos pueden ser PDFs, archivos de texto o incluso bases de datos SQL.

Importante: Debemos asegurar que el texto se extraiga de forma limpia y sin caracteres extraños.

PASO 3: FRAGMENTACIÓN (CHUNKING)

Un LLM tiene un límite de memoria (ventana de contexto). Por eso, dividimos los documentos largos en fragmentos más pequeños (por ejemplo, de 500 a 1000 palabras) que sean manejables para el sistema.

PASO 4: GENERACIÓN DE EMBEDDINGS

Aquí es donde ocurre la magia. Convertimos cada fragmento de texto en una serie de números (un vector) que representa su significado semántico. Para esto usamos modelos de "Embeddings" como los de OpenAI o alternativas locales como *SentenceTransformers*.

PASO 5: ALMACENAMIENTO VECTORIAL

Guardamos estos vectores en una base de datos especializada (Vector Store). Esta base nos permite hacer búsquedas por "similitud semántica" en lugar de por palabras clave exactas.

PASO 6: EL CICLO DE RECUPERACIÓN (RETRIEVAL)

Cuando el usuario hace una pregunta:

1. La pregunta se convierte también en un vector (embedding).

2. Se buscan en la base de datos los 3 o 5 fragmentos de texto cuyos vectores sean más parecidos al de la pregunta.

PASO 7: GENERACIÓN DE LA RESPUESTA (PROMPT ENGINEERING)

Construimos un "Prompt" especial para el LLM que diga algo como:

"Eres un asistente experto. Utiliza el siguiente contexto para responder la pregunta del usuario. Si la respuesta no está en el contexto, di que no lo sabes. No inventes nada."

Contexto: [Fragmentos recuperados]

Pregunta: [Consulta del usuario]

PASO 8: IMPLEMENTACIÓN DEL CÓDIGO

Finalmente, ensamblamos todo en un script de Python que ejecute este ciclo de forma automática cada vez que recibamos una consulta.