

Proyecto: Sistema RAG Local con LlamaIndex y Hugging Face

Descripción Técnica

Este proyecto implementa un sistema de **Generación Aumentada por Recuperación (RAG)** ejecutado de manera local. La solución ingesta una base de conocimiento especializada, compuesta por 5 documentos técnicos, la indexa semánticamente y utiliza un Modelo de Lenguaje (LLM) Open Source para responder consultas basándose estrictamente en la evidencia documental.

La arquitectura se diseñó para operar sobre infraestructura acelerada por GPU (T4) en Google Colab, eliminando la dependencia de APIs externas o servicios de pago.

Arquitectura del Pipeline

El flujo de procesamiento de datos sigue el estándar RAG implementado con el framework **LlamaIndex**:

1. Ingestión de Datos (Data Ingestion)

- **Fuente:** Wikipedia API.
- **Volumen:** 5 Documentos técnicos almacenados localmente en el directorio `mis_documentos_rag/`.
 1. Inteligencia artificial
 2. Red neuronal artificial
 3. Aprendizaje profundo
 4. Procesamiento de lenguaje natural
 5. Visión artificial
- **Persistencia:** Los archivos se descargan en formato `.txt` plano para permitir auditoría manual y validación de contenido.

2. Indexación y Búsqueda (Indexing & Retrieval)

- **Framework:** LlamaIndex (Core).
- **Modelo de Embeddings:** `BAAI/bge-m3`.
 - *Especificación:* Modelo optimizado para representaciones semánticas multilingües y tareas de recuperación densa.
- **Vector Store:** Instancia de `VectorStoreIndex` en memoria RAM para optimizar la latencia durante la fase de recuperación.

3. Generación (Generation)

- **LLM:** `Qwen/Qwen2.5-1.5B-Instruct`.

- *Especificación*: Modelo de 1.5 billones de parámetros con soporte nativo para instrucciones en español.
- **Configuración de Inferencia:**
 - **Temperature: 0.1** (Minimiza la creatividad para priorizar la exactitud factual).
 - **Context Window: 4096 tokens**.
 - **Device Map: Auto** (Asignación automática a GPU).

Estructura de Archivos y Evidencia

La ejecución del proyecto genera los siguientes artefactos verificables:

Plaintext

/content

```
|   └── mis_documentos_rag/    # Base de Conocimiento (5 archivos .txt)
|       ├── Inteligencia_artificial.txt
|       ├── Red_neuronal_artificial.txt
|       └── ...
|
└── evidencia_respuestas.txt # Log de ejecución: Preguntas + Respuestas + Fuentes
└── RAG_Implementation.ipynb # Notebook con el código fuente
```

Validación y Grounding (Fundamentación)

La exactitud de las respuestas se valida mediante el mecanismo de **Source Nodes** (Nodos Fuente) de LlamaIndex. Cada respuesta generada incluye explícitamente:

1. **Texto Sintetizado:** La respuesta generada por el LLM.
2. **Trazabilidad:** Metadatos del documento origen (**filename** y **title**) utilizado para construir dicha respuesta.