RECUPERACIÓN DE INFORMACIÓN AUMENTADA PARA GRANDES MODELOS DE LENGUAJE DE CÓDIGO ABIERTO

UN CASO DE ESTUDIO EN AYUDAS PÚBLICAS DEL GOBIERNO DE ESPAÑA

Desiderio Martí Alcaraz

Web: https://huggingface.co/spaces/DesiMarti/TFMCienciaDatos

Código Fuente: https://github.com/desimartiout/TFM_DesiMarti

UOC - Máster en Ciencia de Datos (Área 2 - Aula 1)

Tutor: José Luis Iglesias Allones

AyudaMe.ai

Problema a resolver

- **Problema para encontrar información sobre ayudas públicas**: La búsqueda de información sobre ayudas y subvenciones públicas es compleja y poco eficiente, debido a la dispersión de datos en portales oficiales y el uso de lenguaje técnico poco accesible a muchos usuarios.
- **Necesidad de sistema que mediante leguaje natural nos permita encontrar ayudas y subvenciones**: Empresas, ciudadanos y gestores necesitan una herramienta que facilite el acceso rápido y comprensible a esta información, sin depender de búsquedas complejas.
- **Desafíos principales**:
  - Extraer datos dinámicos, estructurados y no estructurados de fuentes de datos públicas.
  - Procesar y entender la pregunta de búsqueda del usuario.
  - Garantizar precisión en las respuestas.
  - Implementación con tecnologías OpenSource.

Metodología propuesta

- Implementar una aplicación web (RAG + LLM) que permita la búsqueda de ayudas mediante consultas en lenguaje natural.
- Proyecto con 5 fases principales:
  - Ingesta de datos.
  - Almacenamiento de información en base de datos vectorial.
  - Integración de base de datos vectorial con LLM.
  - Aplicación web.
  - Evaluación del sistema.

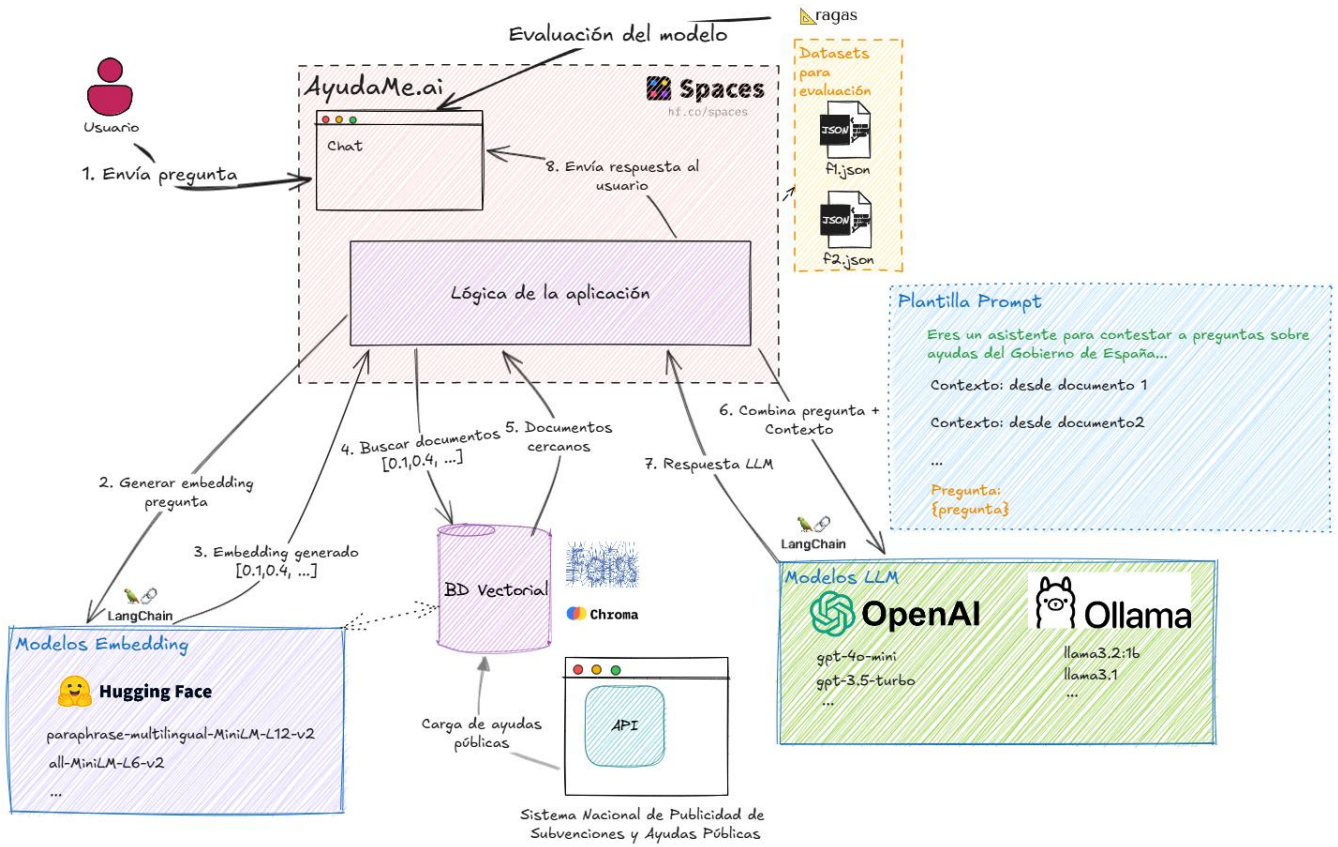
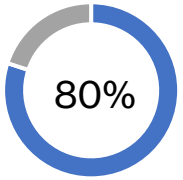


Diagrama arquitectural

Resultados obtenidos

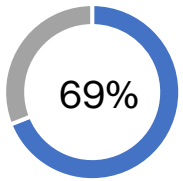
Fortalezas del sistema

- 100% funcional para realizar consultas en lenguaje natural con respuestas simples y comprensibles.
- Experiencia fluida y eficiente gracias a la integración de RAG y modelos LLM.



Similitud semántica con pregunta

Similitud semántica con el texto de referencia



Fidelidad al contexto

Fidelidad de las respuestas al contexto proporcionado

Debilidades

- Existen inconsistencias en muchas respuestas dadas por el sistema, principalmente por la información proporcionada en el contexto.

Áreas de Mejora

- Es necesario mejorar la capacidad para recuperar contexto relevante de la base de datos vectorial y así garantizar una mayor precisión en las respuestas.