PROPUESTA DE PROYECTO FINAL

Nombre y apellido del autor: Fernando Agustin Barzola
Nombre del curso: Generación de Prompts en IA
Nombre del proyecto: Sistema RAG para la Generación Automática de Contexto en Consultas Complejas

---

Presentación del problema a abordar

El problema que abordaré es la dificultad para obtener contextos precisos y relevantes a partir de grandes volúmenes de información. En muchas aplicaciones, como servicios de atención al cliente, educación o investigación, es crucial disponer de respuestas contextuales que integren información precisa y específica. Sin embargo, los sistemas actuales a menudo enfrentan problemas de relevancia y precisión cuando se les consulta sobre temas complejos o amplios. Implementar un sistema RAG que utilice prompts optimizados para generar contexto relevante es relevante porque mejora la precisión y utilidad de las respuestas generadas, ofreciendo resultados más completos y ajustados a la consulta.

Desarrollo de la propuesta de solución

La solución propuesta es desarrollar un sistema RAG que emplee la generación de prompts en dos etapas: recuperación y generación. En la etapa de recuperación, un modelo de búsqueda seleccionará los documentos más relevantes de una base de datos extensa. Posteriormente, en la etapa de generación, un modelo de texto-texto creará respuestas contextuales basadas en la información recuperada. Este sistema se complementará con un modelo de texto-imagen para generar representaciones visuales de los conceptos o contextos generados.

Prompt de ejemplo:

- **Texto-texto (para generación de contexto)**: “Usa la información de los documentos recuperados para responder a la pregunta: ¿Cuáles son los beneficios del aprendizaje profundo en la automatización de tareas de visión por computadora?”

Este prompt se optimizará para que el sistema pueda generar contextos detallados que incluyan las ideas principales y ejemplos relevantes de los documentos recuperados.

Viabilidad del proyecto

El proyecto es técnicamente viable, ya que se basa en el uso de tecnologías accesibles como modelos de recuperación de información (por ejemplo, OpenSearch o FAISS) y modelos generativos de texto (GPT, Llama) . La generación y optimización de prompts permitirá un ajuste eficiente del sistema para mejorar la precisión del contexto generado. El proyecto puede desarrollarse en etapas, comenzando con la creación de prompts básicos y pruebas con conjuntos de datos limitados, seguido de la integración y pruebas de escalabilidad. Los recursos disponibles en plataformas de código abierto y servicios de nube facilitarán la implementación y prueba de la solución.