WICC 2024
XXIII Workshop de Investigadores en Ciencias de la Computación

SISTEMAS DE RECUPERACIÓN AUMENTADA (RAG): UNA PROPUESTA DE INVESTIGACIÓN PARA POTENCIAR LAS BÚSQUEDAS SEMÁNTICAS Y EL CONTEXTO INTERACTUANDO CON INTELIGENCIA ARTIFICIAL GENERATIVA

PONCE, María Paula; MIGO, Gabriel Alejandro; ISTVAN, Romina
Universidad Tecnológica Nacional, Facultad Regional La Plata
Grupo de I&D LINES-IA
{mpaulaponce; gmigo; ristvan}@frlp.utn.edu.ar

Palabras claves: Sistemas de Recuperación Aumentada (RAG), Búsqueda Semántica, Contexto, Inteligencia Artificial Generativa, Grandes Modelos de Lenguaje (LLM), Base de datos Vectoriales

CONTEXTO

Los modelos de lenguaje a gran escala (LLM), han demostrado un rendimiento sobresaliente en una variedad de tareas de procesamiento de lenguaje natural (PLN), incluyendo generación de texto, traducción automática y respuesta a preguntas.

Sin embargo, estos modelos enfrentan desafíos significativos en términos de eficiencia computacional y manejo de información externa. La arquitectura RAG (Sistema de Recuperación Aumentada) emerge como una solución con gran potencial para abordar estas limitaciones.

Este proyecto de investigación se enfoca en mejorar la eficiencia y precisión de la recuperación de información basada en texto mediante el uso de RAG, los cuales combinan técnicas de recuperación de información vectoriales con modelos de lenguaje avanzados para mejorar la relevancia de los resultados de búsqueda. El objetivo es explorar diferentes enfoques y técnicas dentro de los RAG para abrir nuevas oportunidades en la búsqueda semántica y el descubrimiento de conocimiento

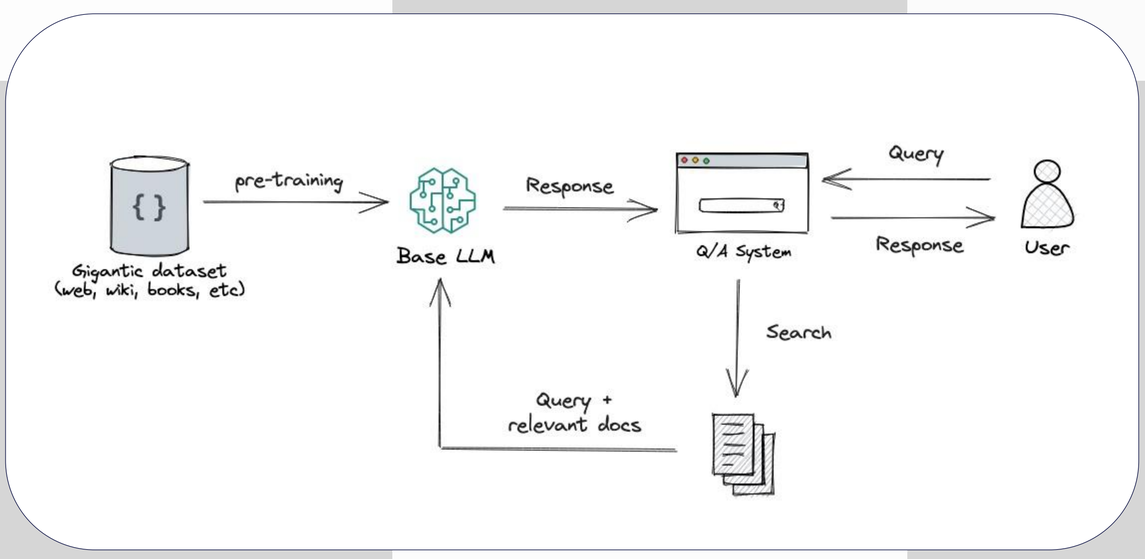

Figura 1: Flujo de información en un sistena RAG entre un Usuario y un LLM

LÍNEAS DE I/D

La línea de investigación propuesta se centra en el desarrollo y la aplicación de sistemas RAG en el contexto de las bases de datos vectoriales para la búsqueda semántica y su integración con grandes modelos de lenguaje en el ámbito de la inteligencia artificial generativa. Esta línea de investigación aborda el desafío de mejorar la generación de respuestas contextualmente adecuadas y la toma de decisiones informadas, evitando problemas ya conocidos de los LLM como las alucinaciones.

RESULTADOS OBTENIDOS / ESPERADOS

Se espera que este proyecto genere avances en la mejora de la respuesta obtenida por los grandes modelos de lenguajes, a través de pasarle un contexto obtenido con los métodos de búsqueda semántica. Estos resultados tendrán implicaciones importantes en diversos campos, desde la investigación en recuperación de información hasta el desarrollo de tecnologías de procesamiento de lenguaje natural más avanzadas. Contribuye a:

- Mejorar la respuesta de LLM.
- Mayor Comprensión del Contenido de los Documentos.
- Integración Fluida con LLM.
- Aplicaciones en la Inteligencia Artificial Generativa.
- Contribuciones a la Investigación en Búsqueda Semántica

FORMACIÓN DE RECURSOS HUMANOS

El equipo del proyecto está formado por docentes investigadores del Grupo de I&D sobre Inteligencia Artificial del laboratorio LINES de la Universidad Tecnológica Nacional Facultad Regional La Plata, un investigador de apoyo, dos tesistas de magister y alumnos becarios de investigación.