Universidad
Carlos III de Madrid

Sistema de verificación de hechos

Alejandro Climent Peñalver – 100539395

Jorge Lázaro Ruiz – 100452172

Aimar Nicuesa Usandizaga – 100537352

Daniel Obreo Sanz – 100451058

Procesamiento de Lenguaje Natural

Máster en Inteligencia Artificial Aplicada

Universidad Carlos III de Madrid

Curso 2024 – 2025

Enlace al repositorio en GitHub

- https://github.com/JorgeyGari/pln_pf.git

Decisiones de diseño

En esta sección se comentan las decisiones tomadas respecto al diseño del chatbot.

Corpus

El corpus empleado es Wikipedia, la enciclopedia libre. Para realizar las búsquedas de información, se emplea la pregunta o enunciado que se está intentando determinar si es verdadero o falso como guía. A partir de este enunciado, se emplea la API de Wikipedia para obtener las páginas más relevantes. Una vez se conocen las páginas, se obtienen las secciones de las mismas que más información pueden proporcionar, para lo que se usa el modelo Llama, que recibirá el nombre de las secciones y el enunciado que se está verificando y seleccionará las mejores secciones de las páginas para obtener la información necesaria que determinará si es verdadera o falsa la afirmación.

Tecnologías

Para implementar el bot se ha usado LangChain, con Llama 3.1 y 3.2 como LLM. Se emplea una cadena de razonamiento, donde el modelo genera una respuesta razonada en base a la información recuperada de Wikipedia para luego evaluar su propio razonamiento. La cadena comienza con la búsqueda en Wikipedia y la extracción de las secciones importantes, sigue con un razonamiento inicial, basado en el contexto y finaliza indicando si su propio razonamiento es válido para determinar si la afirmación es verdadera o falsa, respondiendo además en el idioma en el que se le haya preguntado.

Para el proceso de búsqueda de fuentes de interés se utiliza la versión en inglés de Wikipedia debido a su mayor extensión, por lo que la entrada se traduce automáticamente si se detecta que está escrita en otro idioma. La traducción de la pregunta o enunciado en caso de que este no estuviera en inglés, se utiliza la librería *SpaCy* vista en clase.

Las cadenas secundarias de evaluación de relevancia de las secciones y evaluación de confianza utilizan el formato de salida estructurada de Ollama, que fuerza al modelo de lenguaje a dar una respuesta en un formato JSON preestablecido.

Funcionalidades adicionales

Como funcionalidades adicionales se ha añadido la posibilidad de generar resúmenes basados en el contexto extraído de Wikipedia empleado para indicar si la afirmación era verdadera o falsa, y también se ha añadido la posibilidad de preguntar en diferentes idiomas, y recibir la respuesta en el idioma en el que se ha preguntado. Adicionalmente, el modelo es capaz de evaluar su confianza en la respuesta final proporcionada al usuario.

Evaluación del sistema

Calidad del sistema RAG	El sistema es capaz de recuperar las secciones más importantes de las páginas obtenidas en Wikipedia y razonar la veracidad de una respuesta a partir de las mismas. La recuperación de la información se hace bajo demanda, procesando únicamente las secciones más relevantes de los artículos.
Relevancia y precisión	Los veredictos son precisos a excepción de los momentos donde la pregunta no contiene explícitamente entidades reconocibles dado que esto entorpece la búsqueda en la base de datos; no proporcionando resultados adecuados. Si el enunciado a verificar es lo suficientemente explícito, la respuesta suele ser correcta. Además, la inclusión de una métrica de confianza en la respuesta dada, acompañada de una justificación, hace que los resultados sean explicables e interpretables.
Cobertura de documentos	Relacionado con el apartado anterior, la capacidad de obtener información depende principalmente del enunciado analizado. En la mayoría de ocasiones sí que logra encontrar las evidencias necesarias, pero hay casos en los que puede no obtener resultados si el enunciado no es lo bastante claro.
Tiempo de respuesta	El tiempo de respuesta depende principalmente de la búsqueda en Wikipedia, se sitúa alrededor de 12 segundos, pero depende de la cantidad de información que se busque en la base de datos; y por lo tanto del nivel de detalle del enunciado introducido.

Conclusiones y limitaciones del sistema

El sistema de verificación de hechos desarrollado en esta práctica cumple con los objetivos propuestos, se ha implementado un sistema RAG que permite la extracción de información de una base de datos, en este caso, Wikipedia, y con esa información se determina la veracidad de una pregunta o afirmación.

El sistema es capaz de razonar la respuesta y no tiene demasiadas alucinaciones, pero sí que puede llegar a generar alguna respuesta incorrecta si no dispone del contexto adecuado o si el enunciado es poco claro.