Tarea 04: Agente Conversacional

Curso de Inteligencia Artificial
Escuela de Ingeniería en Computación
Instituto Tecnológico de Costa Rica

I. Objetivo

El propósito de esta tarea es desarrollar una asistente conversacional que pueda desenvolverse con naturaleza ante diferentes preguntas en relación con una base de datos de documentos. De manera que haciendo uso de técnicas de aumentación, y recuperación podamos ampliar el conocimiento de un LLM.

Objetivo principal es realizar una comparación empírica de resultados utilizando dos RAGs con diferente formato de segmentación del texto.

Además, como objetivo secundario, ampliar las capacidades de un LLM conectándolo con herramientas de búsqueda en internet.

II. Pasos a seguir / Instrucciones

A. Base de datos

La base de datos de documentos que utilizaremos serán los apuntes realizados por los estudiantes del curso de Inteligencia Artificial - II Semestre 2025.

La información debe ser extraída de cada uno de los apuntes y agregar la meta data que ustedes consideren necesaria para mejorar los resultados de búsqueda y recuperación del texto, por ejemplo, fecha del documento, nombre del archivo, autor del apunte, etc.

B. Preprocesamiento textual

- Deben de extraer el contenido textual de cada uno de los PDFs
- Aplicar técnicas de limpieza de texto, normalización de caracteres (tildes, minusculas, mayusculas, etc).
- Aplicar técnicas de segmentación de texto (escoger las 2 que ustedes más gusten para comparar)

C. Tokenización y embeddings

Una vez preprocesado el texto deben transformar los datos en embeddings contextuales para ser almacenados en una base de datos vectorial.

El modelo recomendado puede ser *text-embedding-3-small* de OpenAI, sin embargo pueden utilizar cualquiera de su preferencia.

D. Herramientas

Posteriormente debe de crear dos Tools, uno para hacer el RAG (**RAG Tool**), es decir extraer información de la base de datos vectorial y otro que le permita hacer búsquedas en Internet (**WebSearch Tool**).

Para esta parte de la arquitectura vamos a utilizar LangChain, un marco de desarrollo diseñado para construir aplicaciones basadas en modelos de lenguaje.

E. Perfil, orquestación y memoria del agente LLM

En esta etapa se debe de definir el prompt base para otorgar personalizado a su agente, definir su perfil y el rol, incluya un nombre al agente, en qué se especializa, como debe ser el estilo de comunicación y las restricciones y limitaciones al momento de responder.

El Agente debe de responder a preguntas que existan en los apuntes realizados indicando documentos de referencia, y autores, además de la respuesta a las preguntas realizadas por el usuario.

Para la parte de orquestación de las herramientas debe el agente debe decidir autonamente cuál herramienta usar basado en las instrucciones del perfil, las búsquedas en la web solo se deben de permitir cual el usuario explicitamente lo solicita. Se recomienda como Orquestador del agente el modelo gpt-3.5-turbo-0125 por su efectividad y costo económico.

Además incluya una ventana de memoria conversacional temporal que le permita recordar el contexto de la sesión actual. Esto significa que el agente debe ser capaz de mantener coherencia entre preguntas consecutivas (por ejemplo, *¿y cómo se calcula esa derivada?*), sin almacenar conversaciones de manera permanente.

Su Agente debe ser lo más preciso posible y evitar respuestas como “No he encontrado la información” a una pregunta que evidentemente usted sabe que tiene en sus documentos.

F. Aplicación

Este chat debe ejecutarse en la interfaz de un navegador donde se note la interacción entre el Agente y el usuario, para esto se sugiere utilizar herramientas como *streamlit_agent*

G. Entrega

El informe deberá realizarse en LaTeX (Overleaf) utilizando la plantilla IEEE para artículos científicos. Además, se debe adjuntar un **Jupyter Notebook** con el código implementado. La entrega final consistirá en un archivo comprimido (`.zip`) que contenga:

- Código fuente en LaTeX.
- El PDF del informe.
- El Notebook con el código fuente.

TABLE I
RÚBRICA DE EVALUACIÓN PARA EL PROYECTO RAG CON AGENTE LLM.

Criterio	Descripción	Puntos
1. Demostración del RAG (evaluación presencial)	Se evalúa el comportamiento del agente en tiempo real: su capacidad para recuperar información desde la base vectorial, mantener coherencia en las respuestas, utilizar la herramienta adecuada según la intención de la pregunta, y aplicar el contexto conversacional de la sesión. Debe demostrar un funcionamiento correcto de la orquestación y del razonamiento del agente.	50 pts
2. Informe escrito	El documento debe describir de manera clara las etapas de desarrollo del sistema RAG, incluyendo: preprocesamiento de texto, generación de embeddings, indexación, definición del perfil del agente, orquestación de las Tools, uso de memoria conversacional y presentación de resultados.	50 pts
Total		**100 pts**

RÚBRICA

Si el trabajo no se encuentra debidamente ordenado y presentado siguiendo una adecuada estructura para el informe, puede ser considerado como incompleto y cualquiera de las rúbricas se puede ver afectada.