Aplicaciones LLM

Full-stack starter con create-llama

Intro

- “create-llama” es una herramienta desarrollada por LlamaIndex para agilizar la creación de templates de aplicaciones RAG full-stack.

- Limitaciones:
  - No incluye base de datos para alojar los documentos privados. Los documentos privados se incluyen manualmente en un directorio de la app.
  - Por lo tanto, no incluye una funcionalidad CRUD para gestionar los documentos privados.

- Actualmente permite crear tres tipos de aplicaciones full stack:
  - Opción 1: Frontend y Backend con Next.js serverless.
  - Opción 2: Frontend con Next.js y Backend con FastAPI (Python).
  - Opción 3: Frontend con Next.js y Backend con Express (Javascript).

Opción 1: Frontend y Backend con Next.js serverless

- Es la opción más sencilla.
- Solo requiere hacer deployment en Vercel.
- Limitaciones: se quedará corta cuando intentemos escalar la aplicación.

Opción 1: ¿Cómo utiliza LlamaIndex?

- La lógica RAG de LlamaIndex está en:
  - app/api/chat/engine/index.js
- Escrita en Typescript

Opción 2: Frontend con Next.js y Backend con FastAPI

- El tutorial de LlamaIndex sugiere hacer deployment de frontend y backend (ambos) en Render.com

- No me parece buena idea, pues nos limitará a la hora de escalar el frontend de la aplicación. Me parece más adecuado hacer deployment del frontend en Vercel y deployment del backend en Render.com como hicimos con la aplicación ToDo.

Opción 2: ¿Cómo utiliza LlamaIndex?

- La lógica RAG de LlamaIndex está en el folder de backend:
  - backend/app/utils/index.py
- Escrita en Python

Opción 3: Frontend con Next.js y Backend con Express

- Teniendo en cuenta que tanto LlamaIndex, como LangChain como la API de ChatGPT están desarrollados nativamente en Python, creo que lo más recomendable es especializarse en Backend con FastAPI (Python) en lugar de con Express (Javascript).

- Por ese motivo no trataremos esta tercera opción en el curso, si bien puede ser una buena alternativa para desarrolladores que tengan una trayectoria en Javascript o que trabajen en equipos especializados en Javascript.