Aplicaciones LLM

LlamaIndex

Introducción

- En su primera versión, LlamaIndex es una framework menos generalista que LangChain. Quiere hacer menos cosas pero hacerlas mejor. Se especializa en generar posibilidades para aplicaciones RAG profesionales.

- Es una empresa aún muy joven y pequeña. Aún debe encontrar su visión estratégica y su modelo de negocio. Tiene una gran dependencia de la evolución de chatGPT.

- Aunque la primera versión de LlamaIndex no era muy user-friendly, ahora se están esforzando para hacer una segunda versión mejorada en ese sentido.

Quickstart to LlamaIndex

- Load private document
- Create vector database
- Ask questions to the private document
- Save the vector database

Customization options

- parse into smaller chunks
- use a different vector store
- retrieve more context when I query
- use a different LLM
- use a different response mode
- stream the response back

Use cases

- QA
- Chatbot
- Agent
- Structured Data Extraction
- Multimodal

Optimizing

- Advanced Retrieval Strategies
- Evaluation
- Building performant RAG applications for production

Other

- LlamaPacks and Create-llama = LangChain templates
- Very recent, still in beta
- Very interesting: create-llama allows you to create a Vercel app!
- Very interesting: open-source end-to-end project (SEC Insights)
  - llamaindex + react/nextjs (vercel) + fastAPI + render + AWS
  - environment setup: localStack + docker
  - monitoring: sentry
  - load testing: loader.io
  - web: https://www.secinsights.ai/
  - code: https://github.com/run-llama/sec-insights