RAG en la Justicia Mendocina: Más Allá de la Búsqueda

“El desafío no era buscar más rápido, sino entender mejor lo que buscamos.”

Mauricio Ryan
Poder Judicial de Mendoza
mauricioryan@gmail.com

Objetivos: Explicar RAG como una nueva forma de búsqueda "semántica" y compartir casos prácticos

IA generativa - El camino al RAG

- Prompt
- Inyección de contexto (*)
- Definición de RAG (Retrieval-Augmented Generation).
  Es una técnica de inteligencia artificial que combina la **recuperación de información relevante** de una base de datos externa con la generación de respuestas por parte de un LLM, sin necesidad de reentrenamiento.

- Beneficios en el ámbito judicial: Mayor precisión y relevancia en grandes volúmenes de datos.

Ref (*): tamaño del contexto en páginas: grok 500, chatGPT 250, claude 1200, gemini 2500 pero $$$ caro

- Explicación de embeddings

es una **representación vectorial** de datos (como texto, imágenes o incluso audio) en un espacio matemático de múltiples dimensiones, donde se **captura el significado semántico** o las relaciones entre elementos

hay párrafos como estos:

1. “El juez dictó sentencia en un caso de fraude bancario.”
2. “La Corte Suprema revisó la constitucionalidad de una ley ambiental.”
3. “El imputado fue absuelto por falta de pruebas.”

Y vos querés hacerle al sistema una pregunta:

“¿Qué resolvió el juez en el caso del banco?”

ref: chunk size: min ¼ pag, max 2-4 pag, prom 1-2 pag

Crear embeddings de los documentos para la búsqueda semántica

Cada párrafo se transforma en un vector numérico (un embedding).
No hay texto nuevo, solo representaciones matemáticas.
Ahora el sistema tiene un "mapa semántico" del significado de cada párrafo.

[0.21, -0.53, 0.77, ...]
[0.45, 0.90, -0.11, ...]
[-0.62, 0.12, 0.33, ...]

Crear el embedding de la pregunta
El sistema convierte la pregunta "¿Qué resolvió el juez en el caso del banco?" en otro vector, por ejemplo:

[0.23, -0.55, 0.80, ...]

Compara el vector de la pregunta con todos los embeddings de la base.
El más cercano (según la distancia en ese espacio vectorial) es el documento 1

Proyecto 1: Búsqueda Semántica en Sentencias Penales

- Descripción: Trabajo con **5000 sentencias penales** de la Suprema Corte.
- Estrategia: Generar resúmenes de cada sentencia, **crear embeddings de esos resúmenes** y almacenarlos en OpenSearch de AWS.
- Ejemplos de uso: Búsquedas como "ministros en disidencia en temas de violencia de género".
- Resultados y lecciones aprendidas: Mejora en eficiencia y **descubrimiento de patrones**, **choque cultural**.

Proyecto 2: Chat Interactivo con Expedientes Completos

- Desafío: Expedientes grandes que no caben en el contexto de conversación de modelos de IA.
- Estrategia alternativa de RAG: **Embeddings por párrafo en todos los documentos del expediente**, recuperación de los más relevantes para inyectarlos en el contexto.
- Aplicación: Permitir a secretarios judiciales preguntar, extraer datos, detectar diferencias o inconsistencias que requieran revisión.
- Explicación del límite de contexto y cómo RAG lo resuelve.

Ref (*): tamaño del contexto en páginas: grok 500, chatGPT 250, claude 1200, gemini 2500 pero $$$ caro

Conclusiones y Perspectivas Futuras

Resumen de beneficios: RAG como **herramienta transformadora**, más allá de la búsqueda simple, pero impone un **cambio cultural**

¿Cómo imaginan RAG transformando su trabajo diario?

¿preguntas?

mauricioryan@gmail.com