Aplicaciones LLM

Las bases de datos vectoriales

Recuerda: RAG supera context window

- Preparativos:
  - Divide tus datos en pequeños segmentos.
  - Convierte los pequeños segmentos en números (“embeddings”).
  - Carga los embeddings en una base de datos vectorial.

- Ahora cuando preguntas (“query”) al LLM:
  - El LLM va a la base de datos vectorial y busca (“indexing”) solo (“semantic similarity”) los datos que responden a tu pregunta.

Recuerda: Embeddings

- Los ordenadores trabajan con números.
- Por eso convierten el texto en números.
- Hacen lo mismo con las imágenes, el audio, el video, etc.
- Los embeddings son vectores de números.
  - Ejemplo: “hola” se convierte en un embedding como (1,4,6).

Bases de datos vectoriales

- Las bases de datos vectoriales están especializadas en trabajar con cientos de millones de embeddings, son mucho más rápidas que las bases de datos convencionales.

- Optimizadas para:
  - Almacenar (“storing”).
  - Buscar (“indexing”).
  - Recuperar (“retrieving”).

Semantic similarity

- Las bases de datos agrupan los embeddings por su parecido semántico (“semantic similarity”). Por ejemplo, los embeddings de “perro” y “gato” (semánticamente similares al ser ambos animales) se agruparán en la base de datos vectorial.