Aplicaciones LLM

Evaluación: medir los resultados de una app RAG

Evaluación

- ¿Cómo podemos evaluar un sistema RAG?
  - Podemos evaluar cada proceso por separado: proceso de recuperación y proceso de respuesta (síntesis).
  - Y podemos evaluar el sistema completo de principio a fin.

Evaluación aislada del sistema de recuperación

- Evaluación de la calidad de los fragmentos (chunks) recuperados para una consulta (query) del usuario.

- Primero necesitas crear un conjunto de datos de evaluación.
  - Puedes usar conjuntos de datos etiquetados (labelled) por humanos.
  - O feedback de usuarios si tienes la aplicación en producción.
  - O crearlo sintéticamente.

- Ejecuta el recuperador (retriever) en el conjunto de datos (dataset).
  - Entrada: consulta (query).
  - Salida (output): los documentos "ground-truth" relevantes para la consulta, las IDs de las outputs devueltas.

- Mide métricas de clasificación.
  - Tasa de éxito (success rate) / tasa de aciertos (hit rate).
  - MRR (Mean Reciprocal Rank), NDCG (Normalized Discounted Cumulative Gain).
  - Tasa de aciertos (hit rate).

Evaluación de todo el sistema RAG: evaluación E2E

- Evaluación de la calidad del output final para un input concreto.
- Crea un conjunto de datos (dataset).
  - Input: query.
  - (Opcional) output: la respuesta "ground-truth".
- Ejecuta el sistema RAG completo.
- Recopila métricas de evaluación.
  - Si no hay etiquetas (labels): evaluaciones sin etiquetas.
    - Métricas: fidelidad, relevancia, adhesión a guidelines, libre de toxicidad.
  - Si hay etiquetas: evaluaciones con etiquetas.
    - Métricas: correctness, etc.