Trang bìa

Demo Tài liệu Mẫu cho RAG
Tác giả: Nhóm Demo
Ngày: 2025-10-18

Gợi ý sử dụng:
- Dùng file này để test Upload & Reindex
- Đặt câu hỏi: "Tài liệu này nói về điều gì?"
- "Tóm tắt 5 ý chính phần Mở đầu"
- "Định nghĩa 'RAG' trong tài liệu"
- "So sánh phương pháp A và B trong Chương 2"

Mở đầu (Introduction)

Mở đầu: Mục tiêu của tài liệu này là cung cấp một ví dụ ngắn gọn để kiểm thử hệ thống RAG.
Tài liệu bao gồm các phần: định nghĩa, kiến trúc cơ bản, hai phương pháp xử lý (A và B), và một số ghi chú triển khai.
Người đọc có thể sử dụng các câu hỏi thường gặp để xem hệ thống truy xuất đúng ngữ cảnh.
Trong phần này, chúng tôi nhấn mạnh rằng nội dung chỉ nhằm mục đích kiểm thử.

Chương 1: Định nghĩa RAG

Định nghĩa 'RAG' (Retrieval-Augmented Generation):
RAG là một kỹ thuật kết hợp bước truy xuất (retrieval) các mảnh ngữ cảnh liên quan từ kho tri thức với bước sinh (generation) của mô hình ngôn ngữ.
Mục tiêu là tạo câu trả lời có căn cứ (grounded) vào nguồn dữ liệu đã truy xuất, giảm nguy cơ bịa (hallucination).
Các thành phần cốt lõi gồm: bộ tách đoạn (chunker), bộ tạo vector (embedding), kho vector (vector store), và mô-đun sinh trả lời.

Chương 2: Hai phương pháp (A và B)

Phương pháp A (Retrieve-then-Generate):
- B1: Nhúng câu hỏi thành vector và truy xuất top-k chunk theo cosine similarity.
- B2: Xây dựng prompt chỉ sử dụng các chunk truy xuất được.
- B3: Mô hình sinh câu trả lời ngắn gọn, trích nguồn khi cần.
Ưu điểm: Đơn giản, dễ triển khai. Nhược điểm: Nhạy với chất lượng truy xuất ban đầu.

Phương pháp B (Retrieve+Re-rank-then-Generate):
- B1: Truy xuất rộng hơn (oversample).
- B2: Re-rank dựa trên tín hiệu phụ (ví dụ: từ khóa 'mở đầu', 'chương 2', hoặc chính xác thuật ngữ cần định nghĩa).
- B3: Dùng các chunk được xếp hạng lại để sinh trả lời.
Ưu điểm: Tăng độ chính xác khi ngữ cảnh dài. Nhược điểm: Phức tạp hơn một chút.

Chương 3: Hướng dẫn tóm tắt

Gợi ý tóm tắt:
- Với phần Mở đầu, có thể tóm tắt thành 5 ý chính:
  1) Mục tiêu: kiểm thử RAG.
  2) Nội dung: định nghĩa, kiến trúc, phương pháp A/B, ghi chú.
  3) Cách dùng: hỏi về tổng quan, tóm tắt, định nghĩa, so sánh.
  4) Nhấn mạnh: nội dung chỉ để test.
  5) Kỳ vọng: câu trả lời có nguồn #chunkId.
- Khi hỏi 'Tài liệu này nói về điều gì?', nên trả lời tổng quan ngắn gọn về RAG và cách kiểm thử.

Phụ lục: Thuật ngữ & Ghi chú

Thuật ngữ:
- Chunk: Đoạn văn bản nhỏ có độ dài cố định và phần chồng lấn (overlap).
- Embedding: Vector biểu diễn ý nghĩa của văn bản.
- Vector store: Kho lưu trữ các vector để tìm kiếm tương đồng.

Ghi chú:
- Đây là tài liệu mẫu, không nhằm thay thế tài liệu học thuật.
- Khi triển khai thật, cân nhắc lưu vector bền vững (ví dụ: PGVector) và thêm lớp tái xếp hạng.