Slide 1: Khái niệm Cosine Similarity

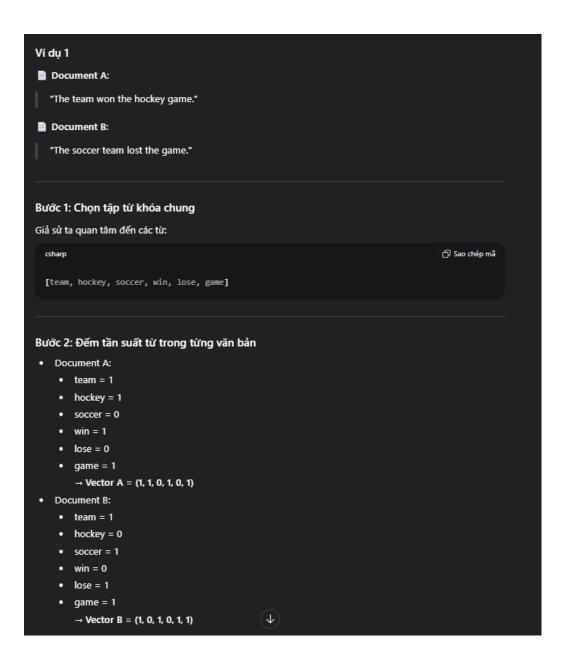
- **Cosine similarity**: đo mức độ giống nhau giữa hai vector trong không gian tích vô hướng.
- Được tính bằng cosine góc giữa 2 vector:
 Tích vô hướng chia tích độ dài

$$sim(x,y) = rac{x \cdot y}{||x|| \; ||y||}$$

- Kết quả
 - = 1 → rất giống (cùng hướng).
 - = 0 → không giống (vuông góc, không chia sẻ thuộc tính).
- Ứng dụng: so sánh văn bản, xếp hạng tài liệu theo từ khóa, phân cụm văn bản, phân tích dữ liệu sinh học.

Slide 2: Úng dụng với Term-Frequency Vector

- Văn bản → vector tần suất từ (term-frequency vector)
 - Mỗi chiều = số lần một từ/phrase xuất hiện.
 - o Dữ liệu thường dài & thưa (nhiều số 0).



Bước 3: Tính cosine similarity

$$sim(A,B) = rac{A \cdot B}{||A|| \; ||B||}$$

Tích vô hướng:

$$A \cdot B = (1 \times 1) + (1 \times 0) + (0 \times 1) + (1 \times 0) + (0 \times 1) + (1 \times 1) = 2$$

• Độ dài A:

$$||A|| = \sqrt{1^2 + 1^2 + 0^2 + 1^2 + 0^2 + 1^2} = \sqrt{4} = 2$$

Độ dài B:

$$||B|| = \sqrt{1^2 + 0^2 + 1^2 + 0^2 + 1^2 + 1^2} = \sqrt{4} = 2$$

· Cosine similarity:

$$sim(A,B)=rac{2}{2 imes 2}=rac{2}{4}=0.5$$

Kết quả: Document A và Document B có mức độ giống nhau = 0.5 (50%).