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

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

- **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: Ứng dụng với Term-Frequency Vector

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

Ví dụ 1

Document A:

> "The team won the hockey game."

Document B:

> "The soccer team lost the game."

Bước 1: Chọn tập từ khóa chung

Giả sử ta quan tâm đến các từ:

```csharp
[team, hockey, soccer, win, lose, game]
```

Sao chép mã

Bước 2: Đếm tần suất từ trong từng văn bản

- Document A:
  - team = 1
  - hockey = 1
  - soccer = 0
  - win = 1
  - lose = 0
  - game = 1

  → **Vector A = (1, 1, 0, 1, 0, 1)**
- Document B:
  - team = 1
  - hockey = 0
  - soccer = 1
  - win = 0
  - lose = 1
  - game = 1

  → **Vector B = (1, 0, 1, 0, 1, 1)**

Bước 3: Tính cosine similarity

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

- **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:**

$$sim(A, B) = \frac{2}{2 \times 2} = \frac{2}{4} = 0.5$$

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