Example Cosine Similarity Between Document Vectors

This document provides an example to explain the concept of calculating the similarity between document vectors using cosine similarity, a common method in text analytics.

# Example Scenario:

Imagine you have a small collection of documents (corpus) related to different animals. We want to determine how similar these documents are to each other based on the words they contain.

## Corpus:

1. Document 1: "The cat sat on the mat."  
2. Document 2: "The dog barked at the cat."  
3. Document 3: "The bird flew over the tree."

# Step 1: Vector Representation of Documents

To compare documents, we first need to represent each document as a vector. Each vector element corresponds to a term's frequency or TF-IDF score in the document.

## Term Frequency Vectors:

Consider the following terms: "cat," "dog," "bird," "sat," "barked," "flew," "mat," "tree."

### Document 1 Vector (TF):

cat: 1/6 ≈ 0.167  
dog: 0/6 = 0  
bird: 0/6 = 0  
sat: 1/6 ≈ 0.167  
barked: 0/6 = 0  
flew: 0/6 = 0  
mat: 1/6 ≈ 0.167  
tree: 0/6 = 0  
  
Document 1 Vector: [0.167, 0, 0, 0.167, 0, 0, 0.167, 0]

### Document 2 Vector (TF):

cat: 1/6 ≈ 0.167  
dog: 1/6 ≈ 0.167  
bird: 0/6 = 0  
sat: 0/6 = 0  
barked: 1/6 ≈ 0.167  
flew: 0/6 = 0  
mat: 0/6 = 0  
tree: 0/6 = 0  
  
Document 2 Vector: [0.167, 0.167, 0, 0, 0.167, 0, 0, 0]

### Document 3 Vector (TF):

cat: 0/7 = 0  
dog: 0/7 = 0  
bird: 1/7 ≈ 0.143  
sat: 0/7 = 0  
barked: 0/7 = 0  
flew: 1/7 ≈ 0.143  
mat: 0/7 = 0  
tree: 1/7 ≈ 0.143  
  
Document 3 Vector: [0, 0, 0.143, 0, 0, 0.143, 0, 0.143]

# Step 2: Cosine Similarity

Cosine similarity measures the cosine of the angle between two vectors. It ranges from -1 (completely dissimilar) to 1 (exactly similar).

The formula for cosine similarity between two vectors A and B is:  
Cosine Similarity = (A · B) / (‖A‖ × ‖B‖)  
Where:  
- A · B is the dot product of the vectors.  
- ‖A‖ and ‖B‖ are the magnitudes of the vectors.

## Calculating Similarity Between Document 1 and Document 2:

1. Dot Product (A · B):  
0.167 × 0.167 + 0 × 0.167 + 0 × 0 + 0.167 × 0 + 0 × 0.167 + 0 × 0 + 0.167 × 0 + 0 × 0 = 0.028

2. Magnitude of Document 1 (‖A‖):  
√((0.167)² + 0² + 0² + (0.167)² + 0² + 0² + (0.167)² + 0²) = √(0.0836) ≈ 0.289

3. Magnitude of Document 2 (‖B‖):  
√((0.167)² + (0.167)² + 0² + 0² + (0.167)² + 0² + 0² + 0²) = √(0.0836) ≈ 0.289

4. Cosine Similarity:  
0.028 / (0.289 × 0.289) ≈ 0.336

So, the cosine similarity between Document 1 and Document 2 is approximately 0.336.

## Calculating Similarity Between Document 1 and Document 3:

1. Dot Product (A · C):  
0.167 × 0 + 0 × 0 + 0 × 0.143 + 0.167 × 0 + 0 × 0 + 0 × 0.143 + 0.167 × 0 + 0 × 0.143 = 0

2. Magnitude of Document 3 (‖C‖):  
√(0² + 0² + (0.143)² + 0² + 0² + (0.143)² + 0² + (0.143)²) = √(0.0614) ≈ 0.248

3. Cosine Similarity:  
0 / (0.289 × 0.248) = 0

So, the cosine similarity between Document 1 and Document 3 is 0, indicating no similarity.

# Interpretation:

Document 1 and Document 2 have a moderate similarity score of 0.336, indicating that they share some common words (e.g., "cat"). Document 1 and Document 3 have a similarity score of 0, indicating no common words and hence no similarity.

# Summary:

Cosine similarity is an effective way to measure how similar two documents are based on the words they contain. In this example, Document 1 and Document 2 are somewhat similar, while Document 1 and Document 3 are completely dissimilar. This method is widely used in text analytics for tasks like document clustering, information retrieval, and plagiarism detection.