Cosine Similarity

2. [8 points] Calculate the cosine similarity between two documents. Cosine similarity is defined as the dot product of the vectors divided by their magnitude.

$cosine\_similarity(A, B) = \cos\theta = \frac{A \cdot B}{||A||||B||}$

where θ is the angle between the two vectors A and B and $A \cdot B$ is the dot product of the two vectors A and B, the dot product can be calculated as $A \cdot B = \sum_{i=1}^{n} A_i B_i = [A_1 B_1 + A_2 B_2 + ... + A_n B_n]$. A_1, A_2, ..A_n and B_1, B_2, ..B_n are the components of the vectors A and B. $||A||$ represents the magnitude of the vectors and can be calculated as $||A|| = \sqrt{A_1^2 + A_2^2 + ...A_n^2}$.

For this assignment, you would need to first calculate the document vectors manually from the document pairs and then calculate the cosine similarity. The document pairs are:

i). *"Today is a nice day. I want to go for a walk."* and *"The man is taking his dog for a walk as the weather is nice today"*. The set of words for this problem is [today, nice, day, want, go, walk, weather, man, dog, take] Consider 'taking' to be the same word as 'take' for constructing the document vector since they are from the same root word.

To calculate the **cosine similarity** between the two documents, follow these steps:

Step 1: Create the document vectors

First, we need to create vectors for the two documents based on the words in the set:

Word set: `[today, nice, day, want, go, walk, weather, man, dog, take]`

Each document will be represented by a vector, where each element corresponds to the frequency of the respective word from the word set in the document.

Document 1: "Today is a nice day. I want to go for a walk."

1. today: 1
2. nice: 1
3. day: 1
4. want: 1
5. go: 1
6. walk: 1
7. weather: 0
8. man: 0
9. dog: 0
10. take: 0 (taking and take are considered the same, but no occurrence here)

So, the vector for Document 1 is:

$$A = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]$$

Document 2: "The man is taking his dog for a walk as the weather is nice today."

1. today: 1
2. nice: 1
3. day: 0
4. want: 0
5. go: 0
6. walk: 1
7. weather: 1
8. man: 1
9. dog: 1
10. take: 1 (since "taking" is considered as "take")

So, the vector for Document 2 is:

$$B = [1, 1, 0, 0, 0, 1, 1, 1, 1, 1]$$

Step 2: Calculate the dot product $A \cdot B$

The dot product is calculated by multiplying corresponding elements of the two vectors and summing the results:

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

$$+(1 \times 1) + (0 \times 1) + (0 \times 1) + (0 \times 1) + (0 \times 1)$$

$$A \cdot B = 1 + 1 + 0 + 0 + 0 + 1 + 0 + 0 + 0 + 0 = 3$$

Step 3: Calculate the magnitude of vectors A and B

The magnitude of a vector A is calculated as:

$$||A|| = \sqrt{A_1^2 + A_2^2 + \cdots + A_n^2}$$

For $A = [1, 1, 1, 1, 1, 1, 0, 0, 0, 0]$:

$$||A|| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2 + 0^2} = \sqrt{6} = 2.449$$

For $B = [1, 1, 0, 0, 0, 1, 1, 1, 1, 1]$:

$$||B|| = \sqrt{1^2 + 1^2 + 0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 1^2 + 1^2 + 1^2} = \sqrt{7} = 2.646$$

Step 4: Calculate cosine similarity

Now, we use the formula for cosine similarity:

$$\text{Cosine Similarity} = \frac{A \cdot B}{||A|| \times ||B||}$$

Substitute the values:

$$\text{Cosine Similarity} = \frac{3}{2.449 \times 2.646} = \frac{3}{6.478} = 0.463$$

Final Result:

The **cosine similarity** between the two documents is approximately **0.463**.

ii). *"The weather is dull today"* and *"How are you doing? Are you going to the college"* Consider the set of words to be [today, weather, dull, go, college, do]. Consider 'going' to be the same word as 'go' for constructing the document vector since they are from the same root word. Similarly, consider 'doing' to be the same word as 'do' for constructing the document vector.

To calculate the cosine similarity between these two documents, we'll first follow the same process as before: construct the document vectors, calculate the dot product, and finally compute the cosine similarity.

Documents:

1. **Document 1**: "The weather is dull today"
2. **Document 2**: "How are you doing? Are you going to the college"

Step 1: Create the document vectors

Word set: `[today, weather, dull, go, college, do]`

Step 1: Create the document vectors

Word set: `[today, weather, dull, go, college, do]`

Document 1: "The weather is dull today"

1. today: 1
2. weather: 1
3. dull: 1
4. go: 0
5. college: 0
6. do: 0 (since "doing" and "do" are considered the same word)

So, the vector for Document 1 is:

$$A = [1, 1, 1, 0, 0, 0]$$

Document 2: "How are you doing? Are you going to the college"

1. today: 0
2. weather: 0
3. dull: 0
4. go: 1 (since "going" is considered as "go")
5. college: 1
6. do: 1 (since "doing" and "do" are considered the same word)

So, the vector for Document 2 is:

$$B = [0, 0, 0, 1, 1, 1]$$

Step 2: Calculate the dot product $A \cdot B$

The dot product is calculated by multiplying corresponding elements of the two vectors and summing the results:

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

$$A \cdot B = 0 + 0 + 0 + 0 + 0 + 0 = 0$$

Step 3: Calculate the magnitude of vectors A and B

For $A = [1, 1, 1, 0, 0, 0]$:

$$||A|| = \sqrt{1^2 + 1^2 + 1^2 + 0^2 + 0^2 + 0^2} = \sqrt{3} = 1.732$$

For $B = [0, 0, 0, 1, 1, 1]$:

$$||B|| = \sqrt{0^2 + 0^2 + 0^2 + 1^2 + 1^2 + 1^2} = \sqrt{3} = 1.732$$

Step 4: Calculate cosine similarity

Now, we use the formula for cosine similarity:

$$\text{Cosine Similarity} = \frac{A \cdot B}{||A|| \times ||B||}$$

Substitute the values:

$$\text{Cosine Similarity} = \frac{0}{1.732 \times 1.732} = \frac{0}{3} = 0$$

Final Result:

The **cosine similarity** between the two documents is **0**.

Since the two documents have no common words from the set, the dot product is 0, and thus the cosine similarity is 0, meaning the documents are orthogonal or dissimilar in terms of the word set provided.