Text Similarity Metric

COSINE SIMILARITY

- Cosine similarity is one of the metric to measure the text-similarity between two documents irrespective of their size in Natural language Processing.
- The Cosine similarity of two documents range from 0 to 1
- Two vectors have the same orientation if Cosine similarity score is 1.
- The value closer to 0 indicates that the two documents have less similarity.

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

EXAMPLE

```
doc_1 = "Data is the oil of the digital economy"
doc_2 = "Data is a new oil"

# Vector representation of the document
doc_1_vector = [1, 1, 1, 1, 0, 1, 1, 2]
doc_2_vector = [1, 0, 0, 1, 1, 0, 1, 0]
```

	data	digital	economy	is	new	of	oil	the
doc_1	1	1	1	1	0	1	1	2
doc_2	1	0	0	1	1	0	1	0

$$\sqrt{\sum_{i=1}^{n} A_i^2} = \sqrt{1+1+1+1+0+1+1+4} = \sqrt{10}$$

$$\sqrt{\sum_{i=1}^{n} B_i^2} = \sqrt{1+0+0+1+1+0+1+0} = \sqrt{4}$$

$$cosine\ similarity = cos\theta = \frac{A \cdot B}{|A||B|} = \frac{3}{\sqrt{10} * \sqrt{4}} = 0.4743$$

$$A \cdot B = \sum_{i=1}^{n} A_i B_i$$

$$= (1*1) + (1*0) + (1*0) + (1*1) + (0*1) + (1*0) + (1*1) + (2*0)$$

$$= 3$$

Cosine Similarity between doc_1 and doc_2 is 0.47

JACCARD SIMILARITY

- Jaccard Similarity defined as an intersection of two documents divided by the union of that two documents that refer to the number of common words over a total number of words
- The Jaccard Similarity score is in a range of 0 to 1. If the two documents are identical, Jaccard Similarity is 1. The Jaccard similarity score is 0 if there are no common words between two documents.

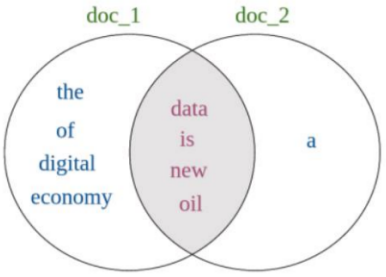
$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}$$

EXAMPLE

$$J(doc_1, doc_2) = \frac{\{'data', 'is', 'the', 'new', 'oil', 'of', 'digital', 'economy'\} \cap \{'data', 'is', 'a', 'new', 'oil'\}}{\{'data', 'is', 'the', 'new', 'oil', 'of', 'digital', 'economy'\} \cup \{'data', 'is', 'a', 'new', 'oil'\}}$$

$$= \frac{\{'data', 'is', 'new', 'oil'\}}{\{'data', 'a', 'of', 'is', 'economy', 'the', 'new', 'digital', 'oil'\}}$$

$$= \frac{4}{9} = 0.444$$



Q3.(Bag of Words & Similarity) **(10 Marks)**

You are provided with the following five sentences.

Sentences:

Sentence 1: The sun rises in the east.

Sentence 2: The sun sets in the west.

Sentence 3: The earth revolves around the sun.

Sentence 4: The moon revolves around the earth.

Sentence 5: The stars are visible at night.

a)Your task is to use the Bag of Words (BoW) model to represent these sentences as vectors.

b)Calculate the **Cosine Similarity** between the following sentence pairs using the BoW vectors:

1. **Sentence 1** and **Sentence 2**
2. **Sentence 1** and **Sentence 5**
3. **Sentence 3** and **Sentence 4**

Bag of Words Vectors:

Sentence 1: 2,1,1,1,1,0,0,0,0,0,0,0,0,0,0,0,0
Sentence 2: 2,1,0,1,0,1,1,0,0,0,0,0,0,0,0,0,0
Sentence 3: 2,1,0,0,0,0,0,1,1,1,0,0,0,0,0,0,0
Sentence 4: 2,0,0,0,0,0,0,1,1,1,1,0,0,0,0,0,0
Sentence 5: 1,0,0,0,0,0,0,0,0,0,0,1,1,1,1,1

b) Cosine Similarity Calculation:

The cosine similarity between two vectors A and B is given by,

$$\text{Cosine Similarity} = \frac{\Sigma(A_i \times B_i)}{\sqrt{\Sigma A_i^2} \times \sqrt{\Sigma B_i^2}}$$

1) Between S1 and S2:

$$\Sigma(S1_i \times S2_i) = (2\times2)+(1\times1)+(1\times0)+(1\times1)+(1\times0)+(0\times1)+(0\times1)+(0\times0)+(0\times0)+(0\times0)+(0\times0)+(0\times0)+(0\times0)+(0\times0)+(0\times0)+(0\times0)$$

$$= 6$$

$$\sqrt{\Sigma S1_i^2} = \sqrt{2^2+1^2+1^2+1^2+1^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2}$$

$$= 2.8284$$

$$\sqrt{\Sigma S2_i^2} = \sqrt{2^2+1^2+0^2+1^2+0^2+1^2+1^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2+0^2}$$

$$= 2.8284$$

$$\text{Cosine Similarity} = \frac{6}{2.8284 \times 2.8284}$$

$$= 0.7500$$

2) Between S1 and S5:

$\Sigma(S1_i \times S5_i) = 2$

$\sqrt{\Sigma S1_i^2} = 2.8284$

$\sqrt{\Sigma S5_i^2} = 2.4495$

$$\text{Cosine Similarity} = \frac{2}{2.8284 \times 2.4495}$$

$$= 0.28868$$

3) Between S3 and S4:

$\Sigma(S3_i \times S4_i) = (2\times2) + (1\times1) + (1\times1) + (1\times1)$

$= 7$

$\sqrt{\Sigma S3_i^2} = \sqrt{2^2+1^2+1^2+1^2+1^2} = 2.8284$

$\sqrt{\Sigma S4_i^2} = \sqrt{2^2+1^2+1^2+1^2+1^2} = 2.8284$

$$\text{Cosine Similarity} = \frac{7}{2.8284 \times 2.8284}$$

$$= 0.8750$$