Bonus Assignment

Cosine Similarity

Example for what you should do:

Calculate the cosine similarity of the two documents given below.

- **Document 1 = 'the best data science course'**
- **Document 2 = 'data science is popular'**

After creating a word table from the documents, the documents can be represented by the following vectors:

	the	best	data	science	course	is	popular
D1	1	1	1	1	1	0	0
D2	0	0	1	1	0	1	1

- $D1 = [1, 1, 1, 1, 1, 0, 0]$
- $D2 = [0, 0, 1, 1, 0, 1, 1]$

Using these two vectors we can calculate cosine similarity. First, we calculate the dot product of the vectors:

$$D1 \cdot D2 = 1 \times 0 + 1 \times 0 + 1 \times 1 + 1 \times 1 + 1 \times 0 + 0 \times 1 + 0 \times 1 = 2$$

Second, we calculate the magnitude of the vectors:

$$\|D1\| = \sqrt{1^2 + 1^2 + 1^2 + 1^2 + 1^2 + 0^2 + 0^2} = \sqrt{5}$$

$$\|D2\| = \sqrt{0^2 + 0^2 + 1^2 + 1^2 + 0^2 + 1^2 + 1^2} = \sqrt{4}$$

Finally, cosine similarity can be calculated by dividing the dot product by the magnitude

$$similarity(D1, D2) = \frac{D1 \cdot D2}{\|D1\|\|D2\|} = \frac{2}{\sqrt{5}\sqrt{4}} = \frac{2}{\sqrt{20}} = 0.44721$$

The angle between the vectors is calculated as:

$$cos(\theta) = 0.44721$$

$$\theta = \arccos(0.44721) = 63.435$$

You are required to measure the cosine similarity between four documents and bring the similarity score sorted descendingly. (More similar to the least similar) (documents of assignment1 that you used before)

Example of the output required

D1 and D2 cosine similarity = 0.9

D1 and D3 cosine similarity = 0.8

D1 and D4 cosine similarity = 0.7

D2 and D3 cosine similarity = 0.6

D2 and D4 cosine similarity = 0.6

D3 and D4 cosine similarity = 0.4