DATA 622 | HOMEWORK 3

Instructions:

1. **Read the File**

Read the content of https://www.washingtonpost.com/world/2025/06/13/air-india-plane-crash-survivor-vishwash-kumar-ramesh/ into a Python variable. Load the first 700 characters.

2. **Split the Text into Sentences**

Split the text into sentences using NLTK's sentence tokenizer.

3. **Load a Pre-trained Embedding Model**

Load a pre-trained sentence embedding model (such as all-MiniLM-L6-v2 from the sentence-transformers library or anyone of your choice).

Secondly, use TF/IDF to vectorize the first ten sentences.

4. **Embed Each Sentence**

Generate an embedding for each sentence.

Print the shape of the embedding for the first sentence.

5. **Compute Similarity Between Sentences**

Compute the cosine similarity between the embeddings of the first and second sentences.

Print the similarity score.