SCS4204/IS4103/CS4104 - Data Analytics

Assignment 1

Download the Zip file in the UGVLE.

The .txt documents in the zip folder are a set of (8) news articles related to three (3) different news topics namely, *Hurricane Gilbert Heads Toward Dominican Coast*, *IRA terrorist attack*, *and McDonald's Opens First Restaurant in China*.

Determine the *Document Similarity*; how similar two or more documents with respect to each other in this document collection.

Use any of the similarity metrics such as cosine similarity, and Euclidean Distance to calculate *Document Similarity.*

Identify the list of .txt documents related to each news topic.

> Ex: Text documents related to the news topics: *Hurricane Gilbert Heads Toward Dominican Coast are* doc 2.txt, doc 3.txt and doc 7.txt

Prepare a PDF file including,

a. Screenshots with an explanation of the tools you used for the above-mentioned *Document Similarity* implementation.
b. Brief explanation of the pre-processing steps you followed.
c. List of .txt documents related to each news topic.
d. Append your full code lines at the end of the PDF file.

Submit a Zip file including the above-prepared **PDF file** and **all code files** you implemented.

Don't just copy-paste someone else's code for the sake of completion. If you are found to commit plagiarism or an act of cheating, you will be penalized depending on the circumstances.

Deadline: Submit on or before 11.55 PM, 14th August 2022 to the UGVLE.