Exercise – 8

Due By: Friday, October 31, 23:55 Hrs

Goals of this exercise:

- To create vector representations of words and documents
- To use the vector representations to mine information and solve useful problems

Data file to be used: **Session-Summary-all-2025-S1.csv**

Process the data file as suggested in the following steps:

1. Read the data into a dataframe. Treat each row (submission) as a ***document***
2. With each document:
    a. Combine the **Topic** and **YourAnalysis** columns to create a unified tex columns
    b. Remove all special characters (use Python library **re**)
    c. Remove **stop words** and **lemmatize** the text (use Python library **nltk**). Understand what is lemmatization, and its importance in text processing.
    d. Store the pre-processed text into a new column of the dataframe
3. Create **vector representations of each document** based on the following methods, and store them in the dataframe:
    a. Count vectorization (use **CountVectorizer** from **sklearn**)
    b. TFIDF vectorization (use **TFIDFVectorizer** from **sklearn**)
    c. Word2vec vectorization (use **Word2vec** from **genism**)
    d. Save the dataframe into a spreadsheet
4. Using each of the above vectorization methods carry out the following:
    a. Calculate pair-wise cosine distance between the documents and visualize / analyze the results.
    b. Calculate pair-wise Euclidean distance between the documents and visualize / analyze the results.
    c. Perform PCA analysis
    d. Create PCA based 2D visualization and analysis
    e. Create 2D t-SNE based visualization and analysis
    f. Using 2D t-SNE coordinates cluster the documents. Analyze the results.
    g. Using the vector representation itself cluster the documents. Analyze the results.
5. Investigate and solve 3 interesting and useful problems, based on the above vector representations, and the information created through the above steps.

Submission Guidelines

- Create a brief document (3-5 pages) summarizing your analysis, results, and learnings.
- Submit the following files to the **E8** submission point on Moodle:
    - The report – PDF document
    - The Python Notebook / source code

***