

ProgDS-2025 Exercise – 9: Text Encoding and Clustering Metrics

Concepts covered:

1. Text encoding
2. Clustering metrics

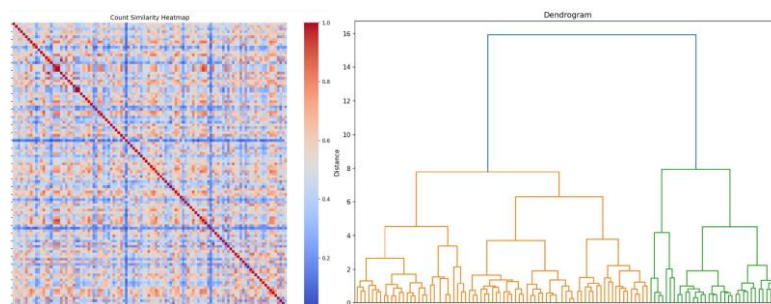
Goals of this exercise:

- To create vector representations of words and documents
- To use the vector representations to group the documents into clusters
- To use metrics like Silhouette Score and Davies-Bouldin Index to decide the optimum cluster count
- To save and analyze clusters, and make conclusions.

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 text column
 - 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 in 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 **gensim**)
 - d. Save the dataframe into a spreadsheet and review all the created data
4. Using **each** of the above vectorization methods carry out the following:
 - a. Calculate pair-wise cosine distance between the documents and visualize (image below) / 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 its analysis
 - e. Create 2D t-SNE based visualization and its analysis
 - f. Using 2D t-SNE coordinates:
 - i. Use Hierarchical Clustering (with complete linkage) to cluster the documents to create 3 to 15 clusters and calculate Silhouette Score and Davies-Bouldin Index values in each case.
 - ii. Create line plots of these metrics and decide the most optimum cluster count.
 - iii. Based on this final count create colour-coded dendrogram (see image below)
 - iv. Against each document record it's cluster number and save the data into a spreadsheet (see below)
 - v. Analyze the clusters and record your observations.
 - g. Using the vector representation itself (ie. without PCA and without t-SNE) cluster the documents, as outlined above, and analyze the results.



Sno	Timestamp	Topic	YourAnalysis	Questions	Comment	Total_Chars	Total_Words	After-Pre-Processing	TF-IDF	word2vec	TFIDF_cluster	Word2vec_cluster
0	2025/10/2	Trade-offs	trade-offs between lazy and eager			1799	240	tradeoff lazy eager exe [0.0. [-5.784796		2		1
1	2025/10/2	Resilient d	resilient data distribution:resilient			893	138	resilient data distribut [0.0.1 [-0.059036		2		1
2	2025/10/2	Aadhaar D	aadhaar data management:the e			985	137	aadhaar data manager [0.508049 [-7.84462]		1		3
3	2025/10/2	Resilient D	resilient distributed dataset:a rei			1326	215	resilient distributed dat [0.0.1 [-8.85856]		2		1
4	2025/10/2	Data stora	data storage in large storage sys			2110	336	data storage large stor [0.0. [-0.066826		2		1
5	2025/10/2	How aadl	how aadhaar stores fingerprint an			1330	191	aadhaar store fingerprin [0.114719 [-6.536105		1		3
6	2025/10/2	Resilient D	resilient distributed datasets:an			1289	194	resilient distributed dat [0.0. [-0.07884]		2		1
7	2025/10/2	Distribut	distributed data processing with			1893	282	distributed data proces [0.0.1 [-0.08628]		2		1
8	2025/10/2	Analysis o	analysis on spark in cloud compu			1531	246	analysis spark cloud co [0.0. [-7.754016		2		1

oooOOOooo

This file is meant for personal use by aashishktrivedi@gmail.com only.
Sharing or publishing the contents in part or full is liable for legal action.