Jacob Dichter

Professor Ausif Mahmood

CPSC 552

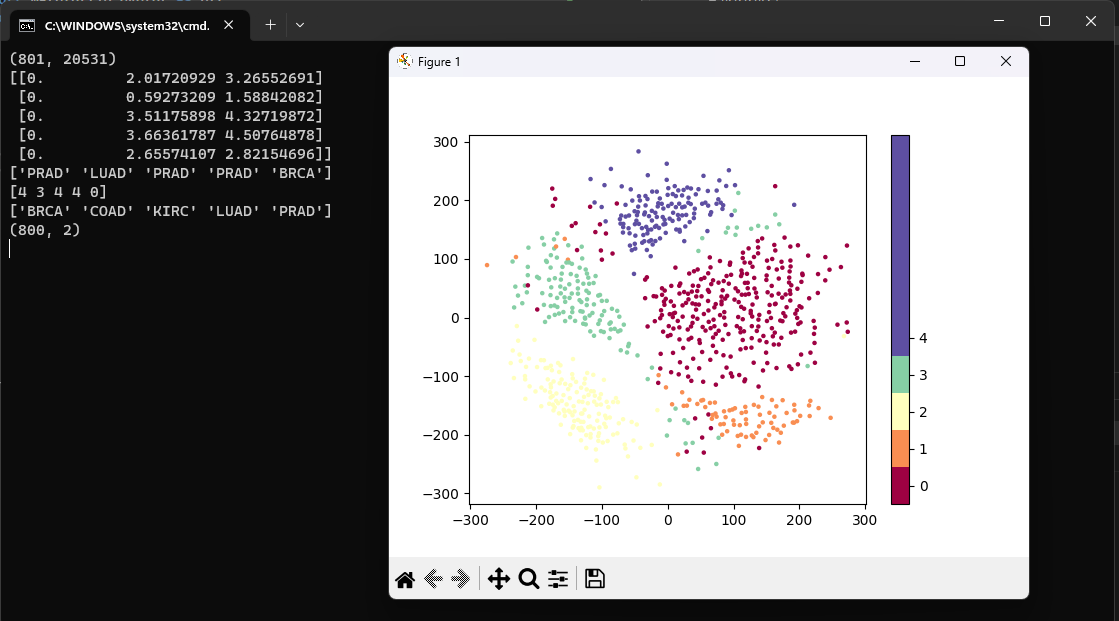
3/26/23

Assignment 8

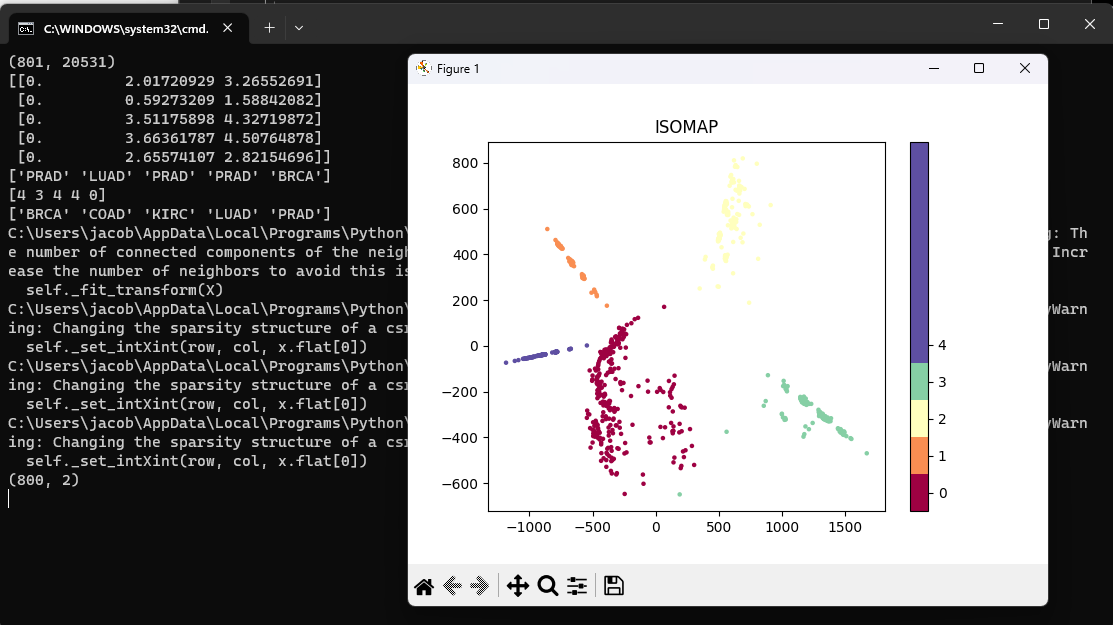
*Data Visualization using PCA, t-SNE and UMap*

It appears that the MDS function does not have the normalized\_stress parameter in the latest version of scikit-learn. This parameter was available in an earlier version but has since been removed. I therefore removed the second keyword argument from MDS().

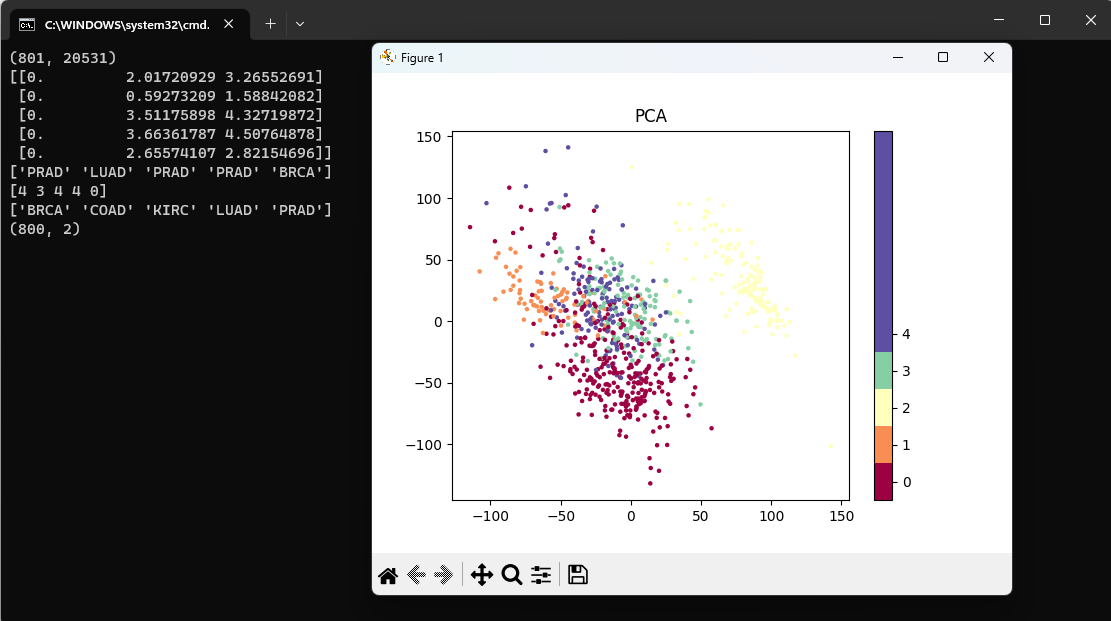
**MDS**



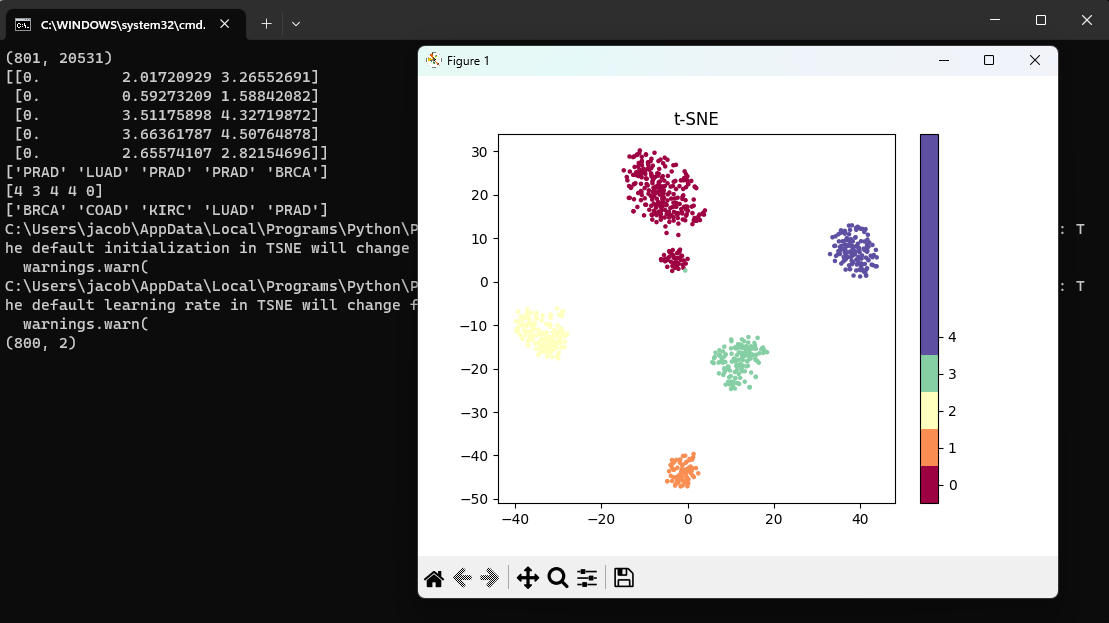
**ISOMAP**

****

**PCA**

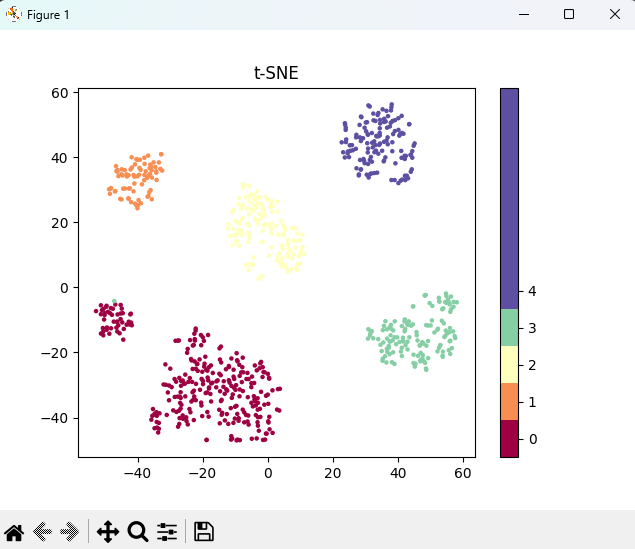
****

**t-SNE**

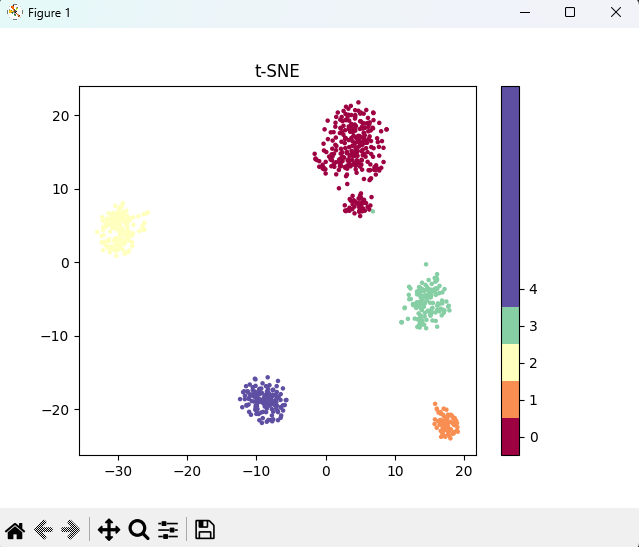
****

***Adjusting Levels of Perplexity for t-SNE***

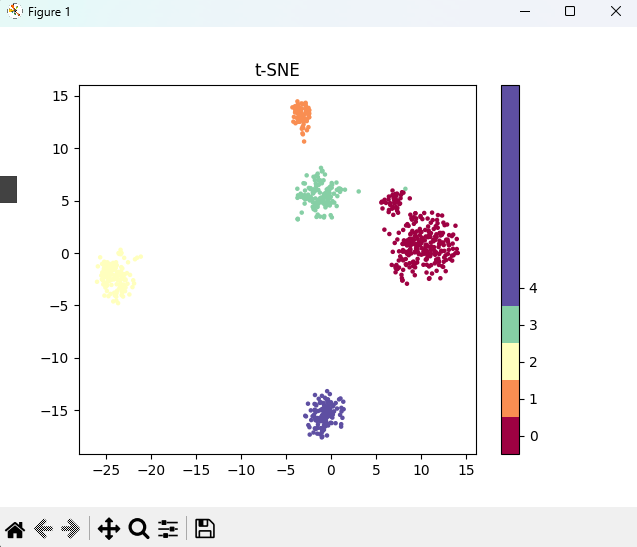
***Perplexity = 10***

******

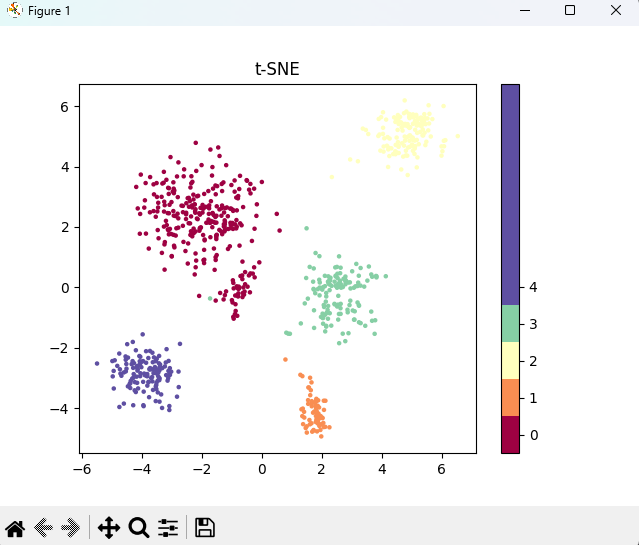
***Perplexity = 60***

******

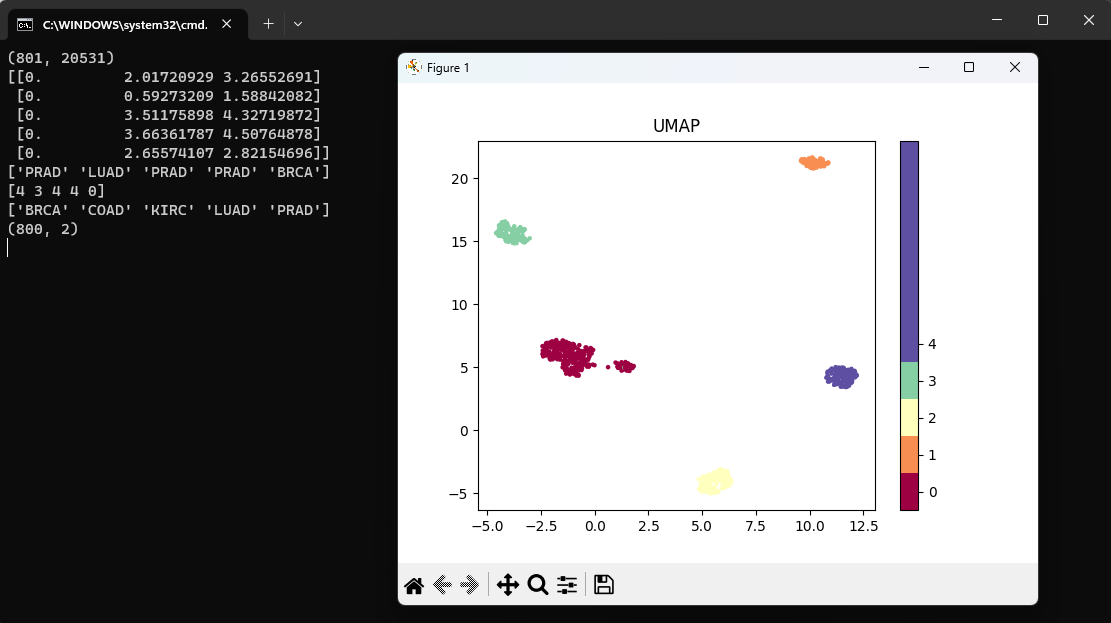
***Perplexity = 100***

******

***Perplexity = 250***

******

**UMAP:**

****

**What are the primary differences and advantages of the techniques used?**

**MDS (Multi-Dimensional Scaling)** is a method that translates pairwise distance information in high dimensional space into 2-D space by using an SVD (Singular Value Decomposition) approach on the distance matrix. The goal of MDS is to preserve the pairwise Euclidean distances in the 2-D space. The main advantage of MDS is that it provides a simple and intuitive representation of the data. However, MDS can be computationally expensive and may not be effective for non-linear relationships in the data.

**ISOMAP (Isometric Mapping)** is based on the MDS approach but preserves geodesic distances between data items. Geodesic distance is the length in terms of the number of edges of the shortest path between two vertices. ISOMAP uses Dijkstra’s algorithm to compute geodesic distance. The advantage of ISOMAP is that it can handle non-linear relationships in the data and can be effective in preserving the global structure of the data. However, ISOMAP can also be computationally expensive and sensitive to the choice of the number of neighbors used in the distance calculation.

**PCA (Principal Component Analysis)** is a linear method that projects high-dimensional data into a lower-dimensional space while maximizing the variance of the projected data. PCA is commonly used for feature selection and data compression. The main advantage of PCA is its computational efficiency and ability to handle high-dimensional data. However, PCA assumes linearity in the data and may not be effective in preserving the non-linear relationships in the data.

**t-SNE (t-Distributed Stochastic Neighbor Embedding)** is a method that models the similarities in the input data space and the embedding space by modeling the distances as distributions and then trying to make the distributions similar by reducing the KL-Divergence between the two as part of the optimization process. t-SNE uses the hyper-parameter “perplexity” to decide how many neighbors will be used to model the distribution in the input data space. The advantage of t-SNE is its effectiveness in preserving the local structure of the data and its ability to handle non-linear relationships. However, t-SNE can be computationally expensive and sensitive to the choice of the perplexity parameter.

**UMAP (Uniform Manifold Approximation and Projection)** is a nonlinear method that is similar to t-SNE. UMAP constructs a high-dimensional graph representation of the data, and then optimizes an equivalent low-dimensional graph to be as structurally similar as possible. To determine connectedness, UMAP extends a radius outwards from each point, connecting points when those radii overlap. UMAP uses two commonly used parameters: n\_neighbors and min\_dist, which control the balance between local and global structure in the final projection. The advantage of UMAP is that it is faster than t-SNE and tends to better preserve the global structure of the data. UMAP also has strong theoretical foundations that allow the algorithm to better strike a balance between emphasizing local versus global structure. However, it is still sensitive to the choice of hyperparameters.

In summary, MDS and PCA are linear methods that work well for data with a linear structure. ISOMAP, t-SNE, and UMAP are nonlinear methods that can handle nonlinear manifolds and preserve the local or global structure of the data. ISOMAP is computationally expensive, t-SNE and UMAP are sensitive to the choice of hyperparameters, and UMAP is faster and tends to better preserve the global structure of the data compared to t-SNE.