CS60050: Machine Learning Mini Project 3

DC-4: Coronavirus Data Clustering using Complete Linkage Hierarchical Clustering Technique

Readme File

Debjoy Saha, 18EC30010

Programming Language used -

- Python

Dependencies -

- Pandas
- Numpy
- Matplotlib
- Random

Steps to Run File -

Change CSV file path(relative to current path) for Data Loading (Line - 440)

```
csv_data_file = 'COVID_4_unlabelled.csv'
```

Set the following parameters - (Lines - 446-448)

```
znorm = True     # If True, using Z-Score Data Normalisation
use_kpp = True   # If True, Initialise centroids using K-means++ algo
n_iters = 20     # No. of iterations to run K-means algorithm
```

In command prompt, run -

```
python clustering.py
```

(clustering.py takes around 30 seconds to 1 minute for complete execution)

Files written to Base Folder -

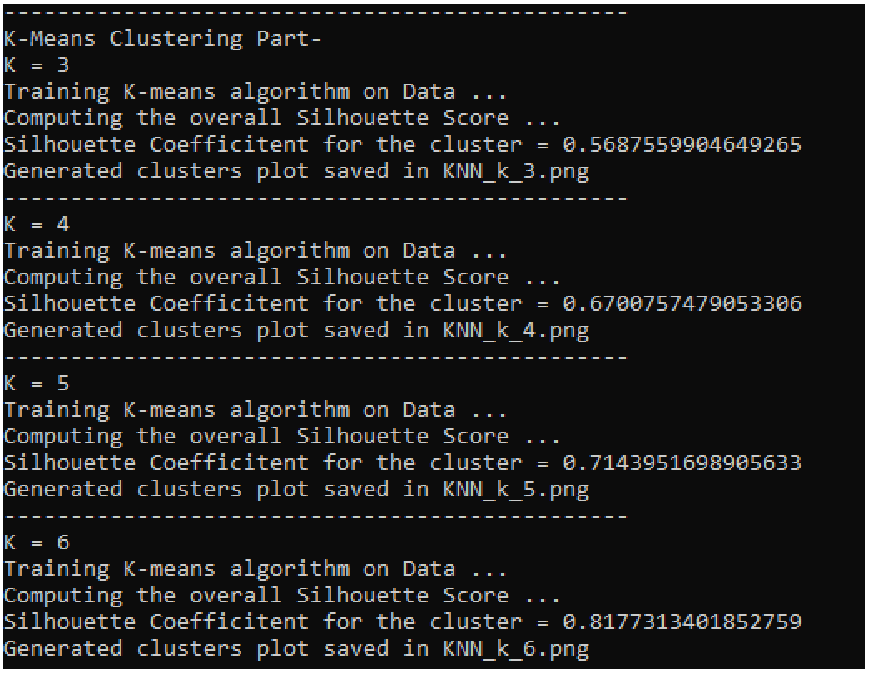
- **KNN clustered Data Plot for k-values - 3, 4, 5, 6**
    - KNN_k_3.png
    - KNN_k_4.png
    - KNN_k_5.png
    - KNN_k_6.png

- **KNN and Agglomerative clustered Data Plot for k value with the best silhouette score**
    - Hierarchical_best_k_{kbest}.png
    - KNN_best_k_{kbest}.png

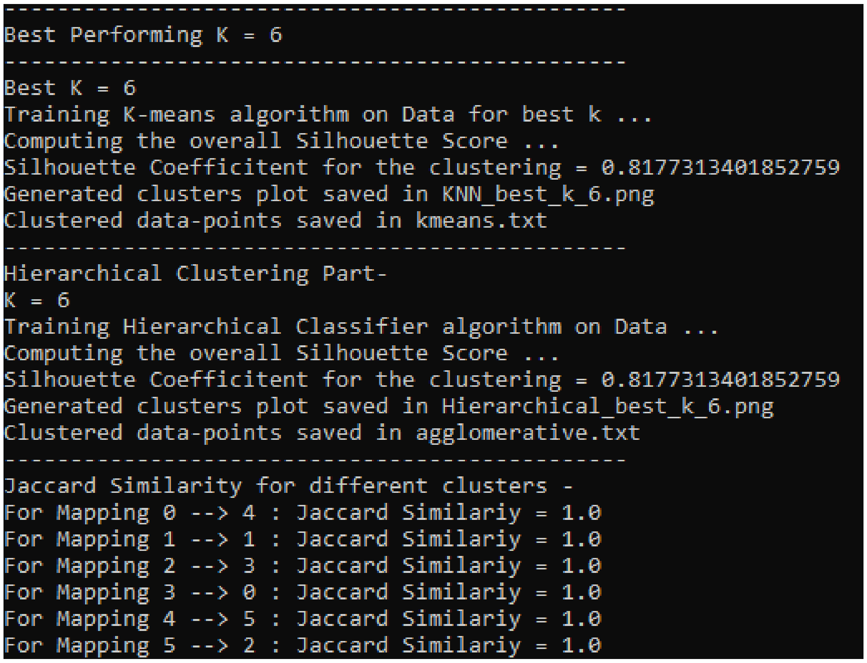
- **Cluster Information corresponding to the best clusters for both K-means and agglomerative clustering in the desired format**
    - agglomerative.txt
    - kmeans.txt

Sample Terminal Output -

Sample terminal output with clustering information, silhouette score for each cluster, and Jaccard Similarity for cluster mappings.



```
-----------------------------------------------
K-Means Clustering Part-
K = 3
Training K-means algorithm on Data ...
Computing the overall Silhouette Score ...
Silhouette Coefficitent for the cluster = 0.5687559904649265
Generated clusters plot saved in KNN_k_3.png
-----------------------------------------------
K = 4
Training K-means algorithm on Data ...
Computing the overall Silhouette Score ...
Silhouette Coefficitent for the cluster = 0.6700757479053306
Generated clusters plot saved in KNN_k_4.png
-----------------------------------------------
K = 5
Training K-means algorithm on Data ...
Computing the overall Silhouette Score ...
Silhouette Coefficitent for the cluster = 0.7143951698905633
Generated clusters plot saved in KNN_k_5.png
-----------------------------------------------
K = 6
Training K-means algorithm on Data ...
Computing the overall Silhouette Score ...
Silhouette Coefficitent for the cluster = 0.8177313401852759
Generated clusters plot saved in KNN_k_6.png
```



```
-----------------------------------------------
Best Performing K = 6
-----------------------------------------------
Best K = 6
Training K-means algorithm on Data for best k ...
Computing the overall Silhouette Score ...
Silhouette Coefficitent for the clustering = 0.8177313401852759
Generated clusters plot saved in KNN_best_k_6.png
Clustered data-points saved in kmeans.txt
-----------------------------------------------
Hierarchical Clustering Part-
K = 6
Training Hierarchical Classifier algorithm on Data ...
Computing the overall Silhouette Score ...
Silhouette Coefficitent for the clustering = 0.8177313401852759
Generated clusters plot saved in Hierarchical_best_k_6.png
Clustered data-points saved in agglomerative.txt
-----------------------------------------------
Jaccard Similarity for different clusters -
For Mapping 0 --> 4 : Jaccard Similariy = 1.0
For Mapping 1 --> 1 : Jaccard Similariy = 1.0
For Mapping 2 --> 3 : Jaccard Similariy = 1.0
For Mapping 3 --> 0 : Jaccard Similariy = 1.0
For Mapping 4 --> 5 : Jaccard Similariy = 1.0
For Mapping 5 --> 2 : Jaccard Similariy = 1.0
```