## REPORT

Understanding Networks through Clustering

B V SOMA ADITHYA

SE21UARI192 | AI-3

1.INTRODUCTION:

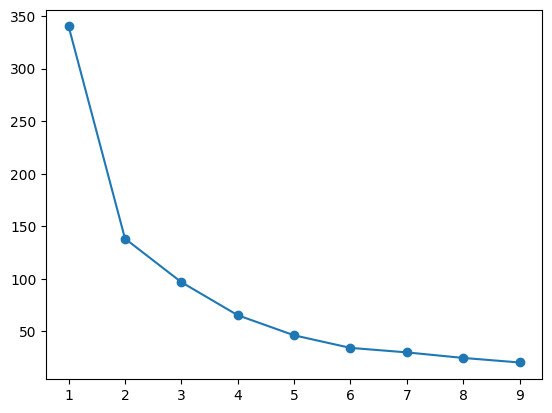
This assignment is on Facebook dataset. It is an undirected network consisting of 4039 nodes and 88234 edges. We need find the no of optimal clusters for this network.

2. Methodology:

For KARATE DATASET.

As karate dataset is provided as an experimental(trail) dataset, I did clustering on the karate dataset first. I used networkx library in this whole procedure. At first I downloaded the .mtx file and read it using mmread function. I drew the graph of karate function and created adjacency matric of it. Then using feature extraction, I extracted 10 features (Eigen vector centrality, betweenness-centrality, degree-centrality, closeness-centrality, Node-degree, page-rank, clustering-coefficient, effective-size, load-centrality**,** average neighbour degree). I used z-score normalization that is standard scaler on data frame for scaling. I created the k-means model with this.

Using elbow graph, I used elbow graph to estimate no of clusters.



I then estimated the silhouette score of the model, got a score of 0.609. As I am successful with this karate data, I used the same code with minor changes for Facebook data. I am attaching the karate dataset Jupyter notebook code file. https://drive.google.com/file/d/1wEozanEONrVtCXcIKtzk37t3uQWRee8W/view?usp=sharing

FOR FACEBOOK DATASET.

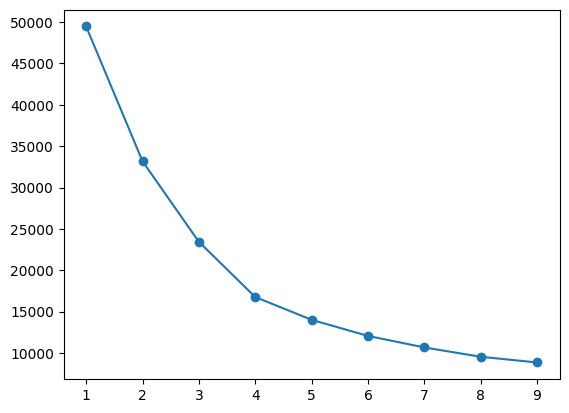
I downloaded the .txt file and read it using using read\_edgelist. I used the same code od karate dataset, I observed that some features are taking much more time, I replaced them with other and a total of 10 features are in the data frame. The 10 features (Eigenvector-centrality, betweenness-centrality, degree-centrality, closeness-centrality, Node-degree, page-rank, clustering-coefficient, subgraph centrality, information-centrality, average-neighbor-degree). I did facebook data with scaling and without scaling too.

* Facebook data with scaling(z-score normalization).

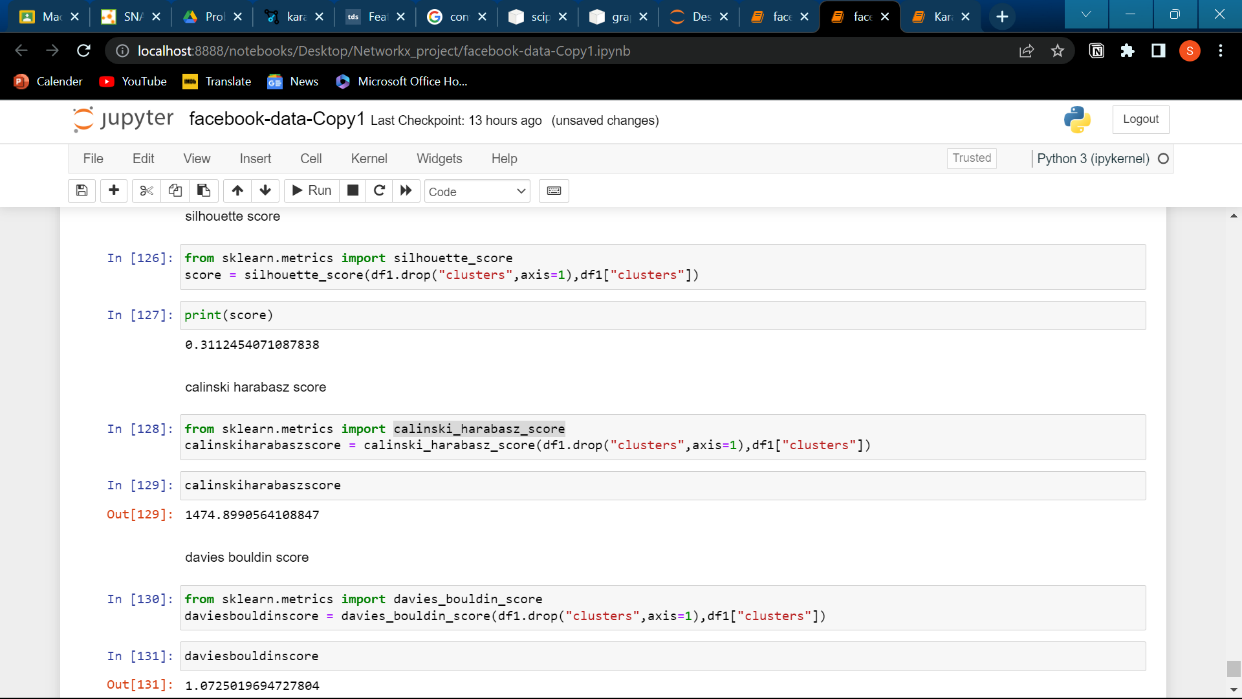
Notebook link: https://drive.google.com/file/d/1sBZOYNY34WJojmeherr7zfCbui6DSBrW/view?usp=sharing

After extracting, I did scaling on data using standard scaler. I created k-means model with this data. And used all 3 evaluation metrices that are silhouette score, calinski harabasz score, davies bouldin score. I used 3,4,5 clusters and the predictions were as follows.

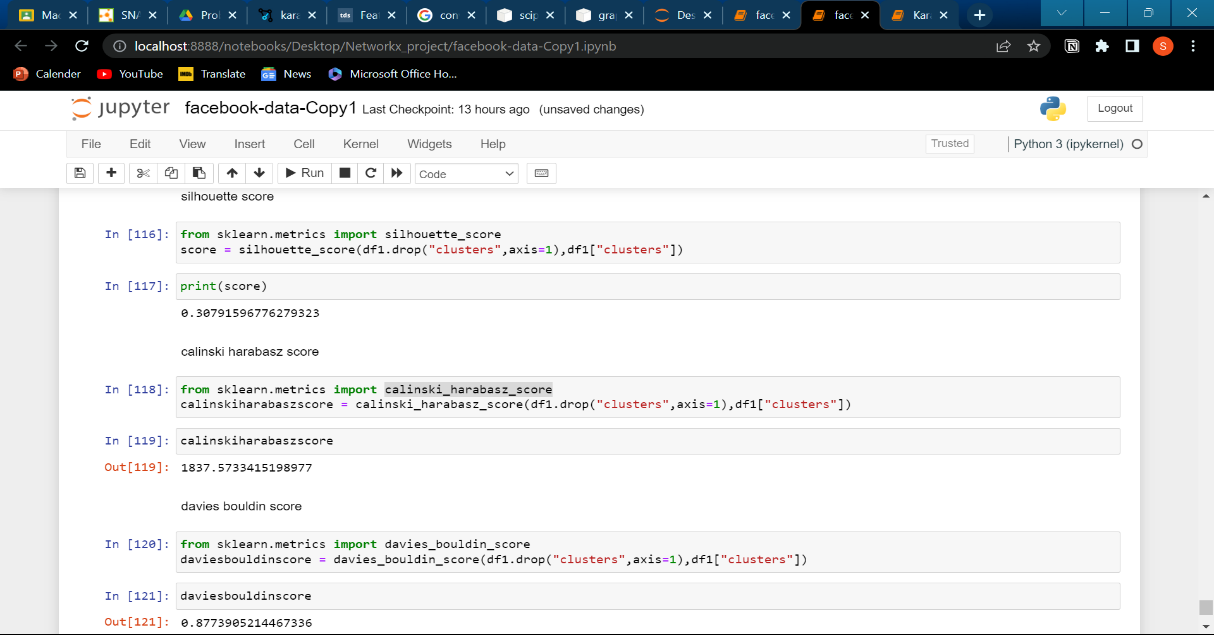
The elbow graph is,



The evaluation scores for 3 clusters are,

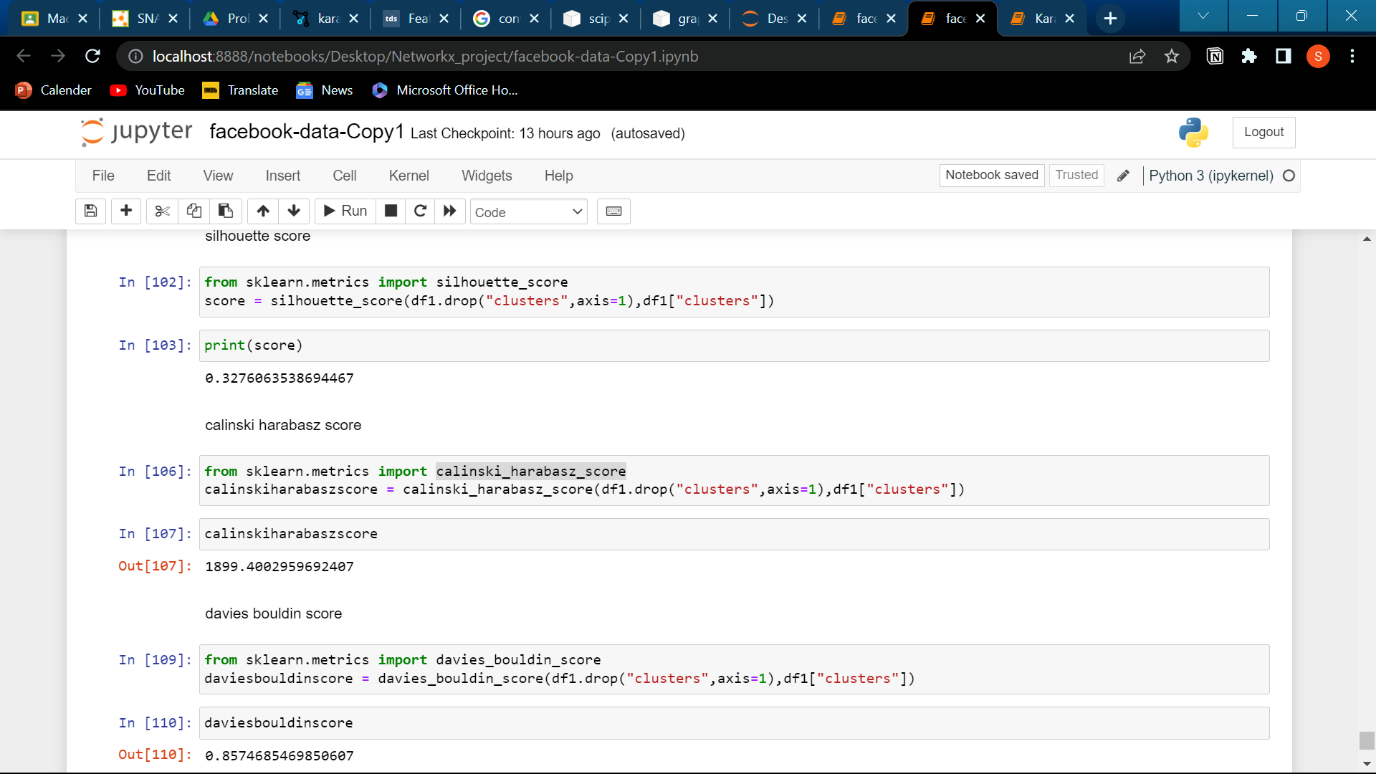


The evaluation scores for 4 clusters are,



|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |

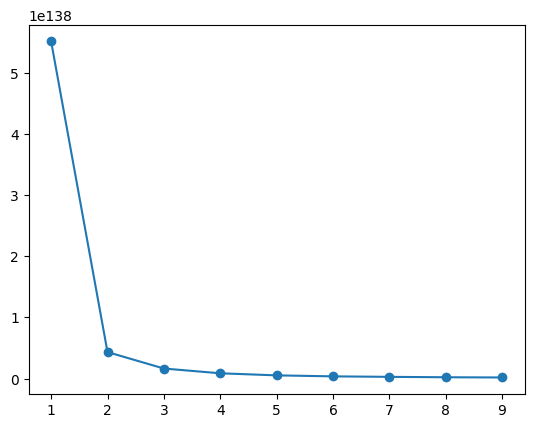
The evaluation scores for 5 clusters are,



* Facebook data without scaling,

With unscaled data I created the k-means model and used the evaluation metric on the model.

With out scaling I got a score 0.978. so I submitted this non scaled model. The elbow graph of this is



3.ANALYSIS:

I brief I extracted 10 features and created a k-means model, and evaluated using silhouette score.

No of optimal clusters are 2.

The silhouette score is 0.978.