Assignment-6 Part-2(Report)

Team Members: 1) Rajan Jhaveri rjj160330
2) Varun Dani vxd162230

Description :

We have performed tweet clustering using k-means algorithm. The measure of distance between two tweets is Jaccard Distance. The program takes initial seeds from Initial-Seeds file provided by the professor. The number of clusters parameter in the program is optional and if not provided will take the default value of 25. We have limited the number of clusters to 25 so that the clustered output makes sense. Best groupings would be between 5-15 clusters because you can get a broader view of the general idea of the tweets as a cluster. The concepts of this assignment can be used in topic mining as well.

Results

The results for various values of k are as follows :

K	SSE
5	141.584336819
10	91.147720736
15	75.9973575725
20	48.8637110896
25	22.5417118337