CS 6375.502

Machine Learning

Submitted By:

Paril Doshi:- PSD170000

Namra Shah:- nxs180003

Part 2

**Tweets Clustering using k-means**

How to compile the code:

The File submitted is a k-mean\_jaccard.py file and made in a editor(Spyder).

The file has no command line outputs. The only output is the “tweets-k-means-output.txt” file which gets generated in same folder.

At the start the tweets are fetched from json file and processed with regular expression and then loaded in “id\_tweet\_dict” dictionary. Then the seed file is loaded in “seed\_list”.

The output file both contains the clustered tweet and the end contains SSE value.

The code is divided into three methods:

1. **Jaccard Distance – Takes two strings as input and outputs the distance.**
2. **Calculate SSE – This function takes cluster and outputs its SSE value.**
3. **K Means – This function takes clusters as k, seed list, Tweet Data file(Dictionary), output.**

The function k\_means is called at the end of file.