CS 6375.003 Machine Learning Assignment 4 Coding Part 2 Report

SHIVAM GUPTA (NET ID: SXG190040)

BHAVYA SREE BOMBAY (NET ID- BXB180036)

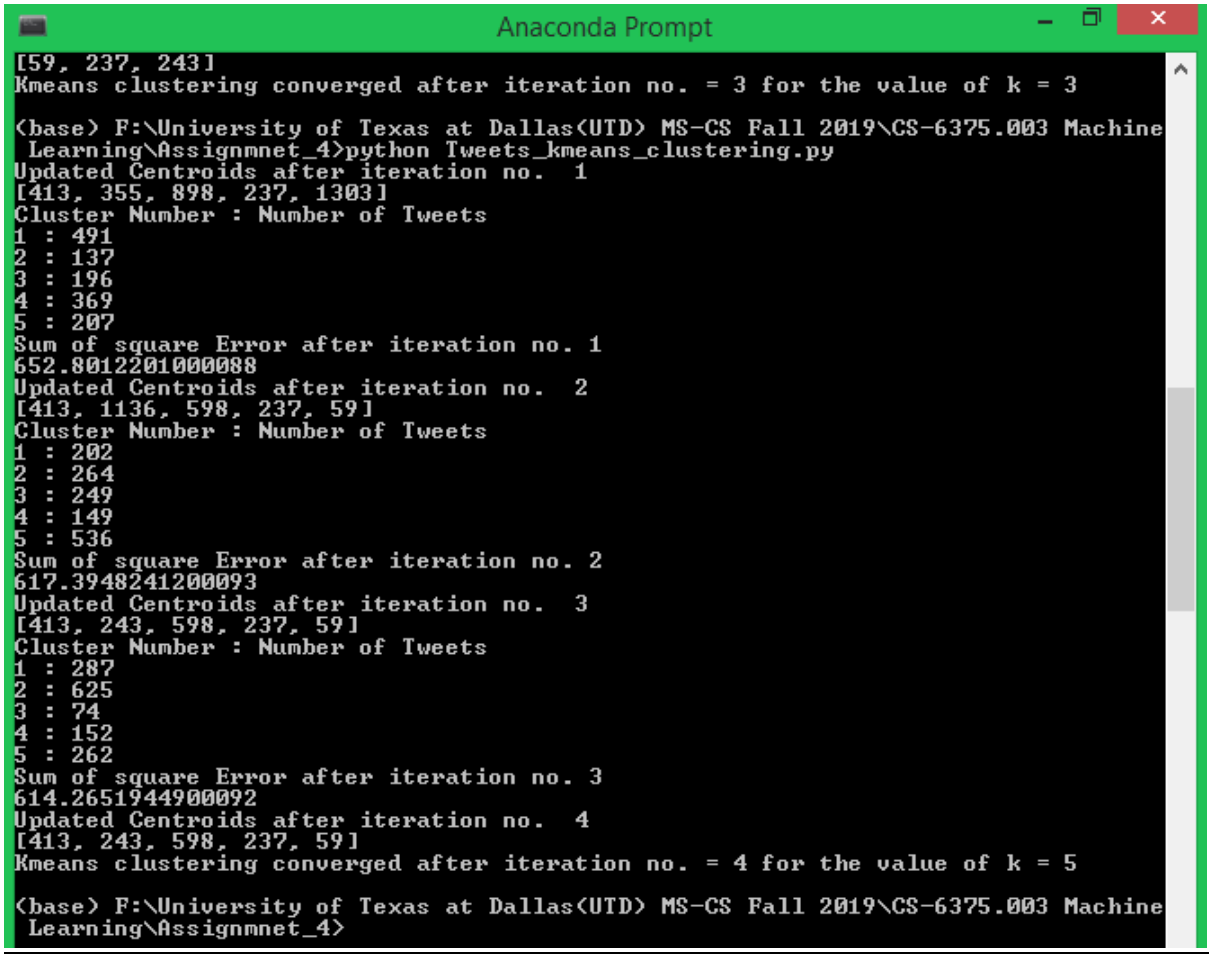
*We have developed k-means clustering from scratch. In the **Pre-processing steps**, we have done the following:*
- Removed the tweet id and timestamp
- Removed any word that starts with the symbol @ e.g. @AnnaMedaris.
- Removed any hashtag symbols e.g. convert #depression to depression.
- Removed any URLs.
- Converted every word to lowercase.

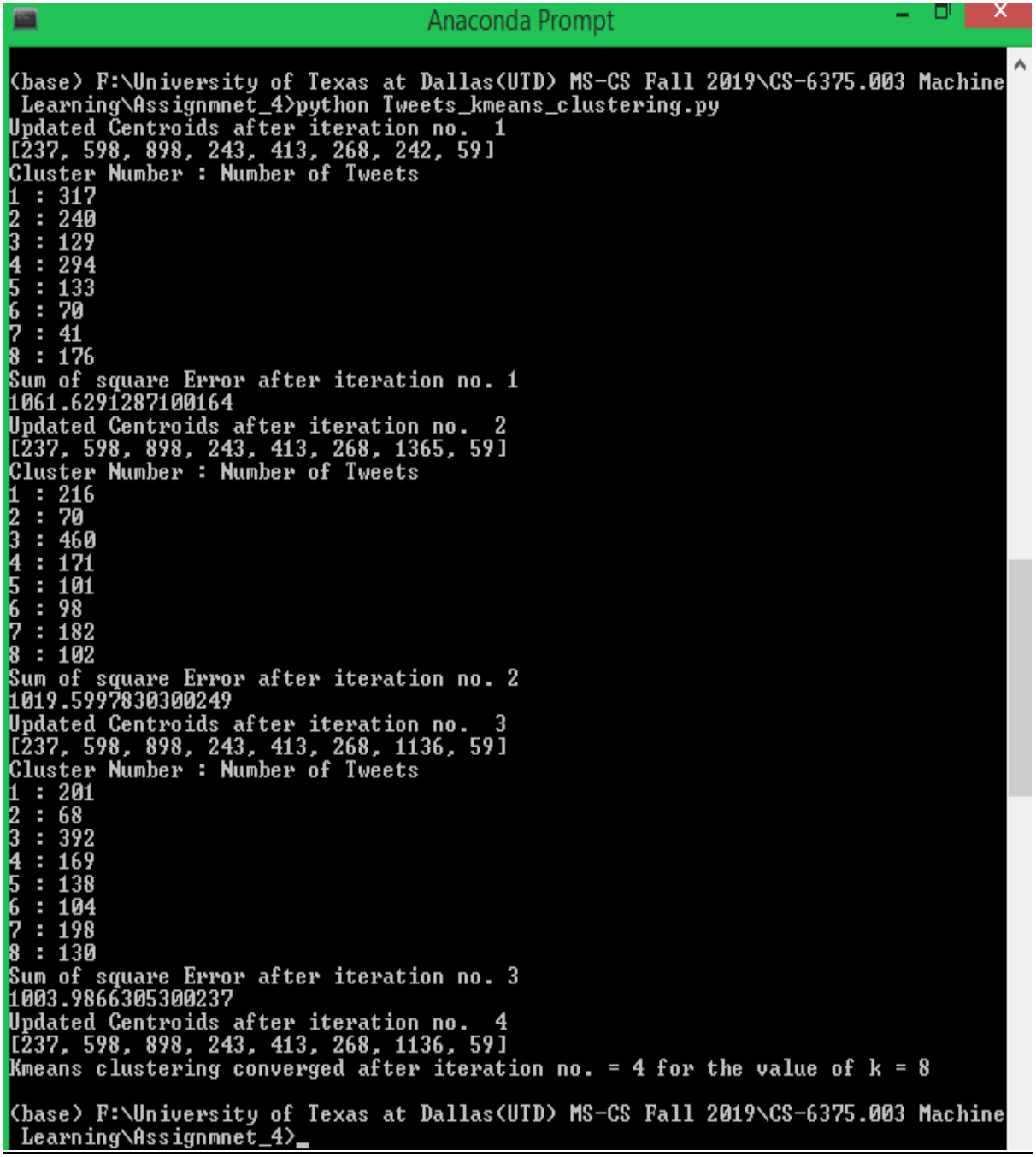
- *The Tweet dataset being used for the following table:*
  foxnewshealth.txt *– The total number of tweets in this dataset: **2000***
- *We have also tested on different tweets text files for eg: msnhealthnews.txt, usnewshealth.txt, bbchealth.txt, etc*
- *The Code can be tested on other datasets by uncommenting line number 114, 115, etc. and change the value of K in line number 117.*
- *We have implemented the convergence criteria when all the centroids remain constant.*

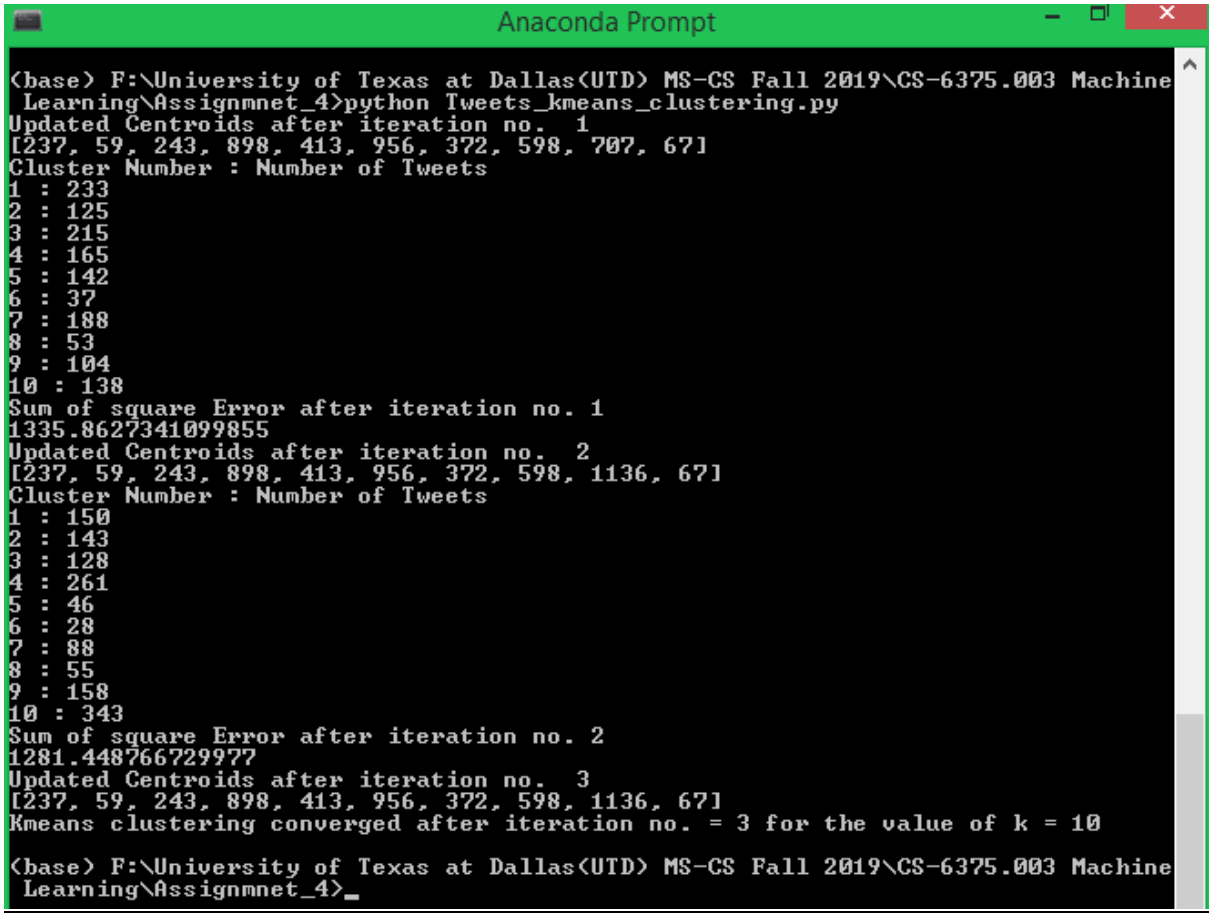
Table showing the Sum of squared error w.r.t Value of K and the size of Cluster

Value of K	Sum of Squared Error (SSE)	Size of Each Cluster (Cluster Number : Number of tweets)
10	1447.7925723000096	1 : 47 2 : 329 3 : 382 4 : 500 5 : 97 6 : 41 7 : 41 8 : 312 9 : 90 10 : 161
8	1157.3646461599678	1 : 431 2 : 71 3 : 88 4 : 216 5 : 491 6 : 358 7 : 240 8 : 105
6	847.2669163699676	1 : 257 2 : 297 3 : 837 4 : 42 5 : 502 6 : 65
5	701.1716522199871	1 : 495 2 : 365 3 : 85 4 : 928 5 : 127
4	563.2013628100026	1 : 107 2 : 571 3 : 354 4 : 968
2	275.94057882000055	1 : 1305 2 : 695

Some of the screenshots of the results for the ushealth.txt tweet dataset:




Anaconda Prompt
(base) F:\University of Texas at Dallas(UTD) MS-CS Fall 2019\CS-6375.003 Machine Learning\Assignmnet_4>python Tweets_kmeans_clustering.py
Updated Centroids after iteration no. 1
[237, 598, 898, 243, 413, 268, 242, 59]
Cluster Number : Number of Tweets
1 : 317
2 : 240
3 : 129
4 : 294
5 : 133
6 : 70
7 : 41
8 : 176
Sum of square Error after iteration no. 1
1061.6291287100164
Updated Centroids after iteration no. 2
[237, 598, 898, 243, 413, 268, 1365, 59]
Cluster Number : Number of Tweets
1 : 216
2 : 70
3 : 460
4 : 171
5 : 101
6 : 98
7 : 182
8 : 102
Sum of square Error after iteration no. 2
1019.5997830300249
Updated Centroids after iteration no. 3
[237, 598, 898, 243, 413, 268, 1136, 59]
Cluster Number : Number of Tweets
1 : 201
2 : 68
3 : 392
4 : 169
5 : 138
6 : 104
7 : 198
8 : 130
Sum of square Error after iteration no. 3
1003.9866305300237
Updated Centroids after iteration no. 4
[237, 598, 898, 243, 413, 268, 1136, 59]
Kmeans clustering converged after iteration no. = 4 for the value of k = 8
(base) F:\University of Texas at Dallas(UTD) MS-CS Fall 2019\CS-6375.003 Machine Learning\Assignmnet_4>



Results and Conclusions:

- *We saw that for different values of k we are getting different values of sum of squared error.*
- *The table above shows the sum of squared errors (SSE) for different values of K and also the size of tweets in each cluster.*
- *In this tweet dataset (**foxnewshealth.txt**) we observed that the sum of squared error got decreased by decreasing the value of K.*
- *There is no specific trend in increment and decrement of the SSE. This trend is going to vary from dataset to dataset and it also depends on the initial seeds (centroids) and we have taken it randomly.*