CS 6375 – ASSIGNMENT 3

Names of students in your group:

Adithya Sundararajan Iyer (asi200000)
Siddhant Suresh Medar (ssm200002)

Number of free late days used: _____2_____

Note: You are allowed a **total** of 4 free late days for the **entire semester**. You can use at most 2 for each assignment. After that, there will be a penalty of 10% for each late day.

Please list clearly all the sources/references that you have used in this assignment.

Programming Part – Part II (70 points)

Dataset Used:

Fox News Health Tweets:
https://raw.githubusercontent.com/siddhantmedar/CS6375-Machine-Learning/main/Tweets/foxnewshealth.txt

Libraries Used:

- math – for general mathematical operations
- random – for generating pseudo-random numbers
- re – for RegEx (regular expression) operation
- copy – for deep copy operations (and shallow, not used)
- urllib – for url handling modules

Data Preprocessing:

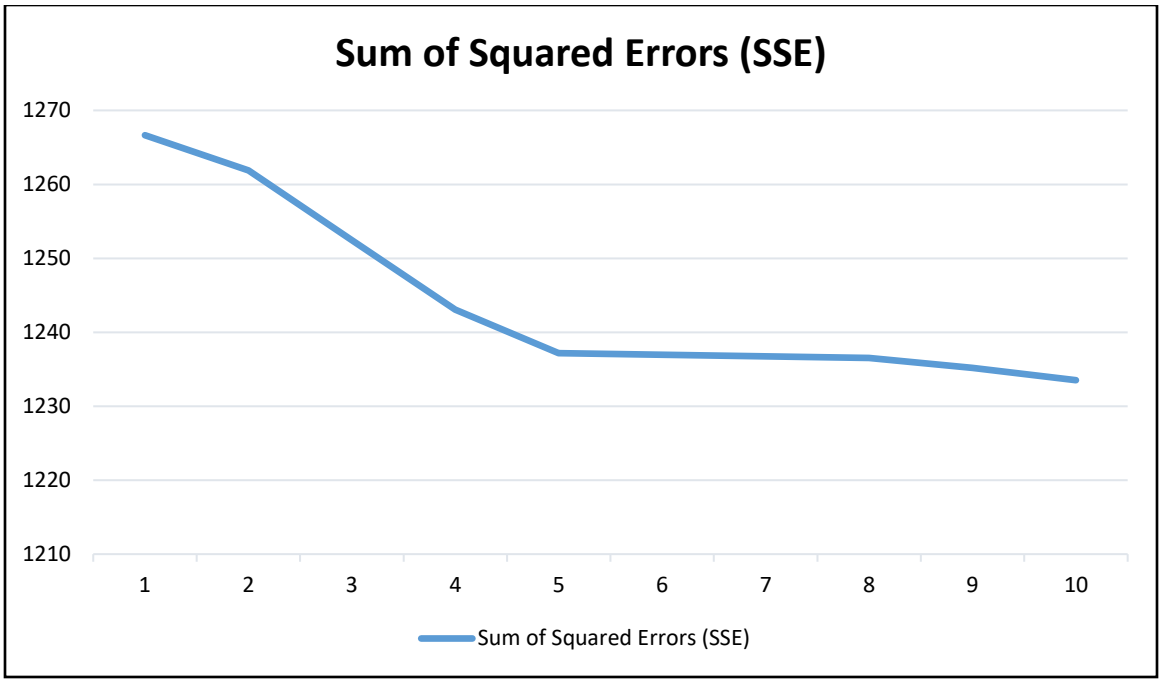
1. Tweet body extracted by removing tweet id and timestamp
2. Twitter handles removed from tweet body (words starting with @)
3. Hashtags replaced with just the word (removed occurrences of #)
4. URLs removed from tweet body (only http links present in dataset)
5. All words converted to lowercase

K-Means Implementation

a) Select a value for hyperparameter **k**
b) Choose *k* unique and random elements as initial centroids
c) Define Jaccard distance to calculate distance between tweets
d) Create *k* clusters and update the centroids
e) Find the sum of squared errors (SSE) after each iteration
f) Repeat until the centroids converge in 2 successive iterations

Results of K-Means Clustering

Value of k	Sum of Squared Errors (SSE)	Size of each cluster
1	1266.6455541800003	1 : 2000
2	1261.8792484100006	1 : 87; 2 : 1913
4	1243.0455140200004	1 : 249; 2 : 1515; 3 : 148; 4 : 88
5	1237.1945713000014	1 : 1395; 2 : 59; 3 : 240; 4 : 219; 5 : 87
8	1236.509608630003	1 : 314; 2 : 53; 3 : 67; 4 : 147 5 : 580; 6 : 361; 7 : 54; 8 : 424
9	1235.185515180001	1 : 860; 2 : 179; 3 : 68; 4 : 203; 5 : 64 6 : 1447 : 181; 8 : 219; 9 : 82
10	1233.518696220004	1 : 205; 2 : 174; 3 : 246; 4 : 106; 5 : 469 6 : 63; 7 : 232; 8 : 89; 9 : 338; 10 : 78



The elbow in the SSE Graph is obtained at k=5. Thus, we see that keeping k=5 is a good trade-off between accuracy and number of clusters(complexity).

References

1. https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter
2. https://docs.python.org/3/library/random.html
3. https://docs.python.org/3/library/re.html
4. https://docs.python.org/3/library/copy.html
5. https://docs.python.org/3/library/urllib.html
6. https://www.statisticshowto.com/jaccard-index/
7. https://365datascience.com/tutorials/statistics-tutorials/sum-squares/