List of sources/references used in this project:

- Link where we obtained the dataset for this assignment (we decided to use the New York Times):

  https://archive.ics.uci.edu/ml/datasets/Health+News+in+Twitter

- Reference used for more details about the Jaccard Distance:

  http://en.wikipedia.org/wiki/Jaccard_index

- Other links used for data processing or clustering understanding:
- https://www.geeksforgeeks.org/python-spilt-a-sentence-into-list-of-words/
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html
- https://scikit-learn.org/stable/auto_examples/cluster/plot_kmeans_silhouette_analysis.html
- https://towardsdatascience.com/visualizing-clusters-with-pythons-matplolib-35ae03d87489
- https://medium.com/@rohithramesh1991/unsupervised-text-clustering-using-natural-language-processing-nlp-1a8bc18b048d
- https://vitalflux.com/k-means-elbow-point-method-sse-inertia-plot-python/#:~:text=Elbow%20method%20requires%20drawing%20a,decreasing%20in%20a%20linear%20fashion)

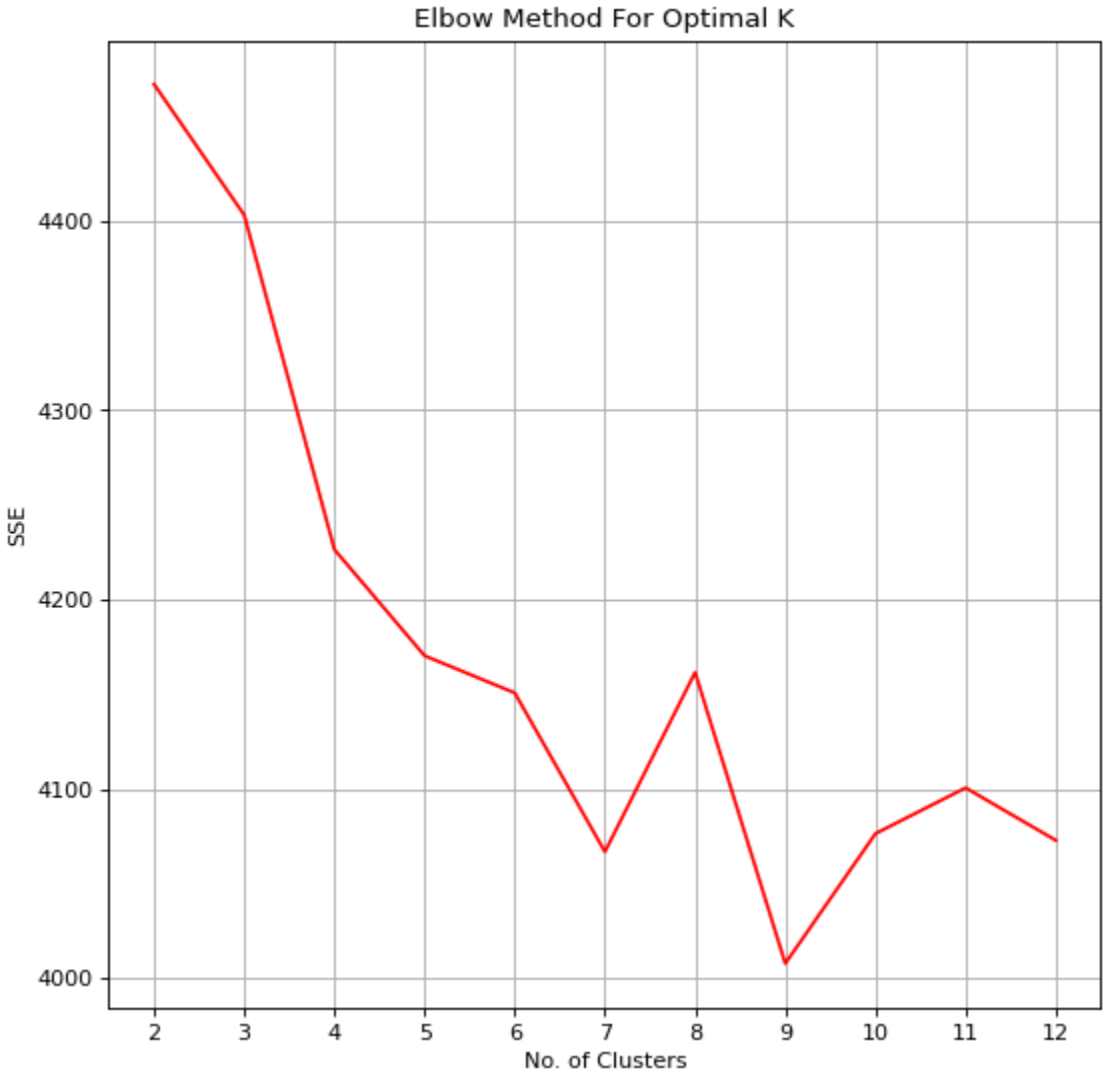
Report File

Log of Trials:

Value of K	SSE	Size of each cluster
2	The SSE is: 4472.30	Cluster 1: 4299 tweets Cluster 2: 1746 tweets
3	The SSE is: 4403.20	Cluster 1: 3250 tweets Cluster 2: 934 tweets Cluster 3: 1861 tweets
4	The SSE is: 4226.52	Cluster 1: 1825 tweets Cluster 2: 2438 tweets Cluster 3: 606 tweets Cluster 4: 1176 tweets
5	The SSE is: 4170.39	Cluster 1: 2316 tweets Cluster 2: 846 tweets Cluster 3: 1203 tweets Cluster 4: 705 tweets Cluster 5: 975 tweets
6	The SSE is: 4150.76	Cluster 1: 395 tweets Cluster 2: 1181 tweets Cluster 3: 1108 tweets Cluster 4: 707 tweets Cluster 5: 1635 tweets Cluster 6: 1019 tweets
7	The SSE is: 4066.83	Cluster 1: 480 tweets Cluster 2: 1353 tweets Cluster 3: 905 tweets Cluster 4: 734 tweets Cluster 5: 244 tweets Cluster 6: 1170 tweets Cluster 7: 1159 tweets
8	The SSE is: 4161.59	Cluster 1: 632 tweets Cluster 2: 36 tweets Cluster 3: 2047 tweets Cluster 4: 998 tweets Cluster 5: 250 tweets Cluster 6: 730 tweets Cluster 7: 362 tweets Cluster 8: 990 tweets

9	The SSE is: 4007.82	Cluster 1: 663 tweets Cluster 2: 921 tweets Cluster 3: 564 tweets Cluster 4: 362 tweets Cluster 5: 621 tweets Cluster 6: 573 tweets Cluster 7: 974 tweets Cluster 8: 953 tweets Cluster 9: 414 tweets
10	The SSE is: 4076.43	Cluster 1: 769 tweets Cluster 2: 290 tweets Cluster 3: 590 tweets Cluster 4: 1055 tweets Cluster 5: 384 tweets Cluster 6: 445 tweets Cluster 7: 602 tweets Cluster 8: 223 tweets Cluster 9: 104 tweets Cluster 10: 1583 tweets
11	The SSE is: 4100.61	Cluster 1: 1845 tweets Cluster 2: 67 tweets Cluster 3: 924 tweets Cluster 4: 1165 tweets Cluster 5: 233 tweets Cluster 6: 552 tweets Cluster 7: 134 tweets Cluster 8: 193 tweets Cluster 9: 386 tweets Cluster 10: 95 tweets Cluster 11: 451 tweets
12	The SSE is: 4072.87	Cluster 1: 609 tweets Cluster 2: 110 tweets Cluster 3: 88 tweets Cluster 4: 407 tweets Cluster 5: 667 tweets Cluster 6: 559 tweets Cluster 7: 1570 tweets Cluster 8: 456 tweets Cluster 9: 405 tweets Cluster 10: 573 tweets Cluster 11: 44 tweets Cluster 12: 557 tweets
Best out of these: 9	**Had SSE of: 4007.82**	

Elbow Method for Optimal K Graph



Analysis of Results:

We can see from the trials and tests of different values of K, that k = 9 yields the lowest Sum of Squared Error (SSE). We can also, see the elbow method graph that was plotted, which shows the SSE vs. the Number of clusters. However, from the graph, we can also see that although the lowest SSE is with k = 9, it could keep decreasing as we keep increasing the value of K (which keeps adding more clusters to separate the tweets into).