Reg Gonzalez CS 6375.001

Assignment 2 - Report

Table of Results

Value of K	SSE (Sum of Squared Errors)	Size of each cluster
5	5020.955678756318	Cluster 1: 758 tweets Cluster 2: 1282 tweets Cluster 3: 601 tweets Cluster 4: 652 tweets Cluster 5: 2952 tweets
6	4999.816159369615	Cluster 1: 1277 tweets Cluster 2: 975 tweets Cluster 3: 1981 tweets Cluster 4: 714 tweets Cluster 5: 337 tweets Cluster 6: 961 tweets
7	4924.548665898326	Cluster 1: 398 tweets Cluster 2: 615 tweets Cluster 3: 1004 tweets Cluster 4: 481 tweets Cluster 5: 558 tweets Cluster 6: 975 tweets Cluster 7: 2214 tweets
8	4857.487008942031	Cluster 1: 559 tweets Cluster 2: 1195 tweets Cluster 3: 1045 tweets Cluster 4: 971 tweets Cluster 5: 766 tweets Cluster 6: 664 tweets Cluster 7: 732 tweets Cluster 8: 313 tweets
9	4920.474091457647	Cluster 1: 1193 tweets Cluster 2: 108 tweets Cluster 3: 468 tweets Cluster 4: 1716 tweets Cluster 5: 882 tweets Cluster 6: 374 tweets Cluster 7: 263 tweets Cluster 8: 1146 tweets Cluster 9: 95 tweets

Assignment 2 - Report

It is important to note that SSE generally decreases as the number of clusters increases. For example, when I tested it with K = 100, 110, 120, 130, and 140, the SSE for 140 was ~4000. This is obviously an improvement to the SSE obtained from the data in the table above. For the sake of brevity, I didn't want to create a table that included the numbers of hundreds of clusters, so that's why I decided to show the results for K = 5, 6, 7, 8, and 9.

Program Output

```
C:\Users\regmc\OneDrive\Desktop\CS 6375\Assignment 2\Assignment2_Part2>python .\Assignment2_Part2.py
K-means for k = 5:
Cluster 1:  758 tweets
Cluster 2:  1282 tweets
Cluster 3:  601 tweets
Cluster 4:  652 tweets
Cluster 5:  2952 tweets
Sum of squared errors: 5020.955678756318


K-means for k = 6:
Cluster 1:  1277 tweets
Cluster 2:  975 tweets
Cluster 3:  1981 tweets
Cluster 4:  714 tweets
Cluster 5:  337 tweets
Cluster 6:  961 tweets
Sum of squared errors: 4999.816159369615


K-means for k = 7:
Cluster 1:  398 tweets
Cluster 2:  615 tweets
Cluster 3:  1004 tweets
Cluster 4:  481 tweets
Cluster 5:  558 tweets
Cluster 6:  975 tweets
Cluster 7:  2214 tweets
Sum of squared errors: 4924.548665898326


K-means for k = 8:
Cluster 1:  559 tweets
Cluster 2:  1195 tweets
Cluster 3:  1045 tweets
Cluster 4:  971 tweets
Cluster 5:  766 tweets
Cluster 6:  664 tweets
Cluster 7:  732 tweets
Cluster 8:  313 tweets
Sum of squared errors: 4857.487008942031


K-means for k = 9:
Cluster 1:  1193 tweets
Cluster 2:  108 tweets
Cluster 3:  468 tweets
Cluster 4:  1716 tweets
Cluster 5:  882 tweets
Cluster 6:  374 tweets
Cluster 7:  263 tweets
Cluster 8:  1146 tweets
Cluster 9:  95 tweets
Sum of squared errors: 4920.474091457647



C:\Users\regmc\OneDrive\Desktop\CS 6375\Assignment 2\Assignment2_Part2>
```