Assignment 3 Report

Group Members:
Kiran Noolvi(Net Id: KXN180017)
Sanjana Annamaneni(Net Id: SXA180095)

Results:

Value of K	SSE	Size of each cluster
5	SSE: 785.0100328776816	1->Cluster Id: 1060, No of Tweets: 131 2->Cluster Id: 1049, No of Tweets: 123 3->Cluster Id: 1327, No of Tweets: 99 4->Cluster Id: 623, No of Tweets: 702 5->Cluster Id: 573, No of Tweets: 345
10	SSE: 715.1147184700712	1->Cluster Id: 552, No of Tweets: 69 2->Cluster Id: 91, No of Tweets: 135 3->Cluster Id: 573, No of Tweets: 175 4->Cluster Id: 949, No of Tweets: 63 5->Cluster Id: 418, No of Tweets: 71 6->Cluster Id: 715, No of Tweets: 129 7->Cluster Id: 623, No of Tweets: 367 8->Cluster Id: 585, No of Tweets: 70 9->Cluster Id: 1310, No of Tweets: 157 10->Cluster Id: 479, No of Tweets: 164
15	SSE: 701.4048021591162	1->Cluster Id: 529, No of Tweets: 43 2->Cluster Id: 842, No of Tweets: 92 3->Cluster Id: 451, No of Tweets: 182 4->Cluster Id: 629, No of Tweets: 148 5->Cluster Id: 958, No of Tweets: 38 6->Cluster Id: 790, No of Tweets: 98 7->Cluster Id: 91, No of Tweets: 84 8->Cluster Id: 573, No of Tweets: 288 9->Cluster Id: 690, No of Tweets: 127 10->Cluster Id: 844, No of Tweets: 22 11->Cluster Id: 597, No of Tweets: 41 12->Cluster Id: 973, No of Tweets: 67 13->Cluster Id: 1368, No of Tweets: 10 14->Cluster Id: 961, No of Tweets: 96 15->Cluster Id: 171, No of Tweets: 64

Value of K	SSE	Size of each cluster
20	SSE: 700.0498019808678	1->Cluster Id: 1064, No of Tweets: 98 2->Cluster Id: 790, No of Tweets: 57 3->Cluster Id: 136, No of Tweets: 28 4->Cluster Id: 91, No of Tweets: 63 5->Cluster Id: 625, No of Tweets: 182 6->Cluster Id: 244, No of Tweets: 76 7->Cluster Id: 1366, No of Tweets: 52 8->Cluster Id: 1143, No of Tweets: 55 9->Cluster Id: 138, No of Tweets: 81 10->Cluster Id: 648, No of Tweets: 32 11->Cluster Id: 477, No of Tweets: 112 12->Cluster Id: 640, No of Tweets: 156 13->Cluster Id: 824, No of Tweets: 47 14->Cluster Id: 531, No of Tweets: 113 15->Cluster Id: 88, No of Tweets: 28 16->Cluster Id: 1327, No of Tweets: 43 17->Cluster Id: 1380, No of Tweets: 35 18->Cluster Id: 1207, No of Tweets: 17 19->Cluster Id: 1082, No of Tweets: 79 20->Cluster Id: 73, No of Tweets: 46
25	SSE: 660.4601432604458	1->Cluster Id: 1348, No of Tweets: 14 2->Cluster Id: 35, No of Tweets: 24 3->Cluster Id: 1251, No of Tweets: 16 4->Cluster Id: 597, No of Tweets: 100 5->Cluster Id: 6, No of Tweets: 36 6->Cluster Id: 618, No of Tweets: 53 7->Cluster Id: 1049, No of Tweets: 104 8->Cluster Id: 1285, No of Tweets: 45 9->Cluster Id: 365, No of Tweets: 76 10->Cluster Id: 318, No of Tweets: 60 11->Cluster Id: 1006, No of Tweets: 42 12->Cluster Id: 533, No of Tweets: 34 13->Cluster Id: 451, No of Tweets: 123 14->Cluster Id: 452, No of Tweets: 49 15->Cluster Id: 703, No of Tweets: 38 16->Cluster Id: 1058, No of Tweets: 16 17->Cluster Id: 1023, No of Tweets: 58 18->Cluster Id: 371, No of Tweets: 22 19->Cluster Id: 573, No of Tweets: 150 20->Cluster Id: 357, No of Tweets: 81 21->Cluster Id: 1327, No of Tweets: 33 22->Cluster Id: 897, No of Tweets: 81 23->Cluster Id: 462, No of Tweets: 26 24->Cluster Id: 649, No of Tweets: 83 25->Cluster Id: 145, No of Tweets: 36

Value of K	SSE	Size of each cluster
30	SSE: 655.5351976357252	1->Cluster Id: 603, No of Tweets: 48 2->Cluster Id: 1067, No of Tweets: 22 3->Cluster Id: 1374, No of Tweets: 25 4->Cluster Id: 53, No of Tweets: 34 5->Cluster Id: 941, No of Tweets: 35 6->Cluster Id: 591, No of Tweets: 1 7->Cluster Id: 682, No of Tweets: 31 8->Cluster Id: 938, No of Tweets: 51 9->Cluster Id: 623, No of Tweets: 217 10->Cluster Id: 798, No of Tweets: 40 11->Cluster Id: 629, No of Tweets: 67 12->Cluster Id: 283, No of Tweets: 20 13->Cluster Id: 587, No of Tweets: 29 14->Cluster Id: 1327, No of Tweets: 39 15->Cluster Id: 466, No of Tweets: 42 16->Cluster Id: 1082, No of Tweets: 57 17->Cluster Id: 597, No of Tweets: 23 18->Cluster Id: 1115, No of Tweets: 30 19->Cluster Id: 212, No of Tweets: 102 20->Cluster Id: 656, No of Tweets: 36 21->Cluster Id: 1364, No of Tweets: 66 22->Cluster Id: 709, No of Tweets: 23 23->Cluster Id: 621, No of Tweets: 43 24->Cluster Id: 239, No of Tweets: 39 25->Cluster Id: 1090, No of Tweets: 26 26->Cluster Id: 731, No of Tweets: 50 27->Cluster Id: 690, No of Tweets: 74 28->Cluster Id: 828, No of Tweets: 17 29->Cluster Id: 696, No of Tweets: 35 30->Cluster Id: 444, No of Tweets: 78

Report:

- The program when executed asks for the value of k.
- The tweets are preprocessed, Then a random initial seed list of side k is created.
- jaccard function returns the distance after taking the input a and b string.
- We use K means clustering to form cluster of tweets and get the final list of centroids.
- Then the SSE is calculated.