Bonus Assignment: k-means Clustering Algorithm

REPORT

Implementation:

Selecting the initial set of centroids:

- ⇨ Centroids are selected randomly with random function.

Stopping condition:

- ⇨ After 5 iterations of k-means clustering, it automatically stops.

Part A:

K-means algorithm:

Implement the k-means algorithm that uses a vector space representation for the documents with tf-idf weighting.

Cluster Size (k)	Average RSS	Time for Computation (Sec)
3	10373.81	24.168
5	10306.18	40.688
4	10305.60	56.007

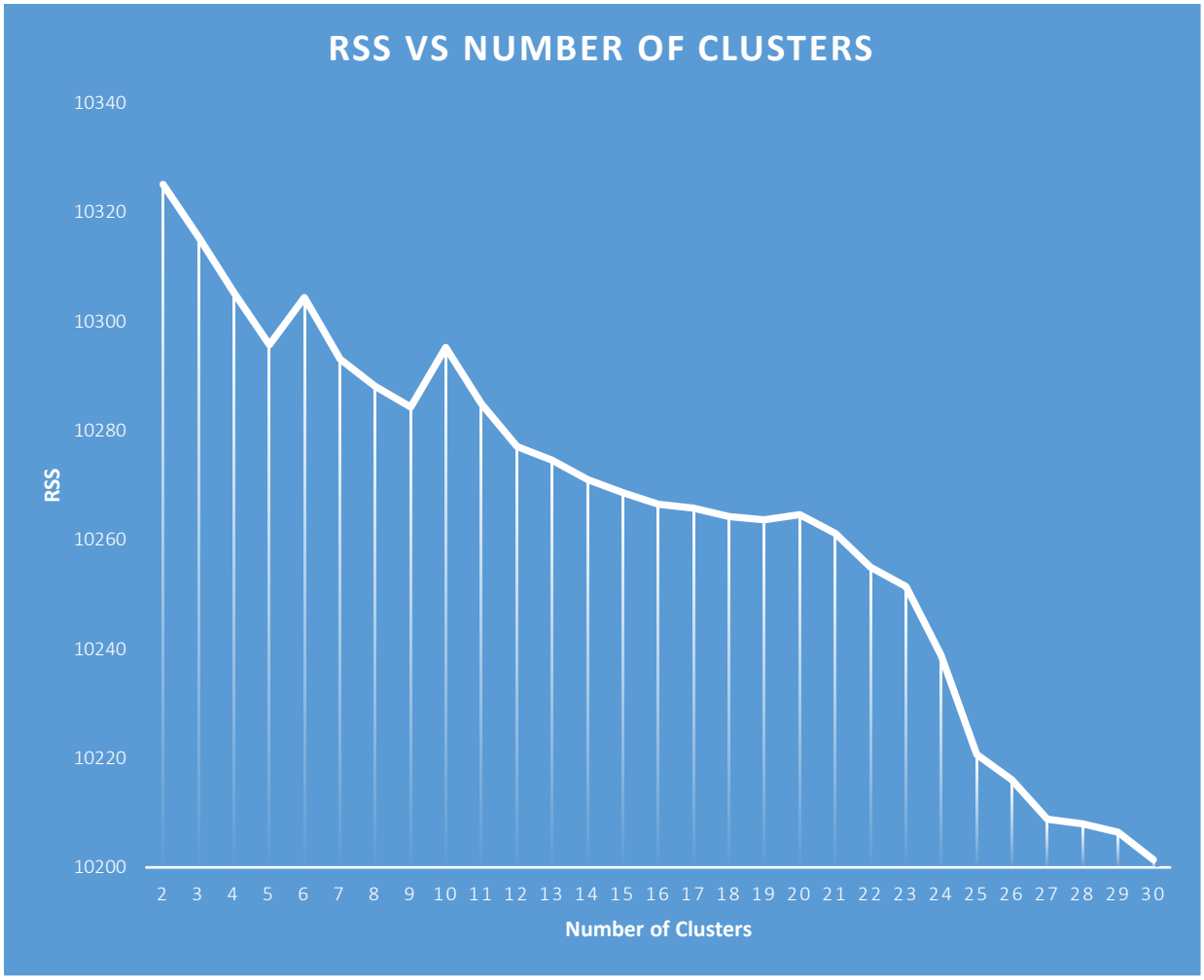
Part B:

Experimental Study:

Conduct an experimental study, with the TIME dataset, to understand the relationship between RSS and the number of clusters (k). The goal is to measure the RSS values for various cluster sizes ranging from k=2 to k=30.

Number of clusters	Avg. RSS
2	10325.08688
3	10315.37828
4	10305.35706
5	10295.65227
6	10304.35706
7	10293.10986
8	10288.11104
9	10284.34111
10	10295.27172
11	10284.87552
12	10277.10697
13	10274.61855
14	10271.06668
15	10268.64752
16	10266.45273
17	10265.83017
18	10264.26256
19	10263.70153
20	10264.65708
21	10261.15689
22	10255.03126
23	10251.38292
24	10238.90478
25	10220.71199
26	10216.00234
27	10208.79319
28	10208.00234
29	10206.39619
30	10201.40498

Plot comparing the RSS values with k -



From the graph, we can say that at 12, 25 and 27, k provides a good tradeoff.

Conclusion:

Final clustering result depends on selection of initial centroid and K- mean algorithm always converges.