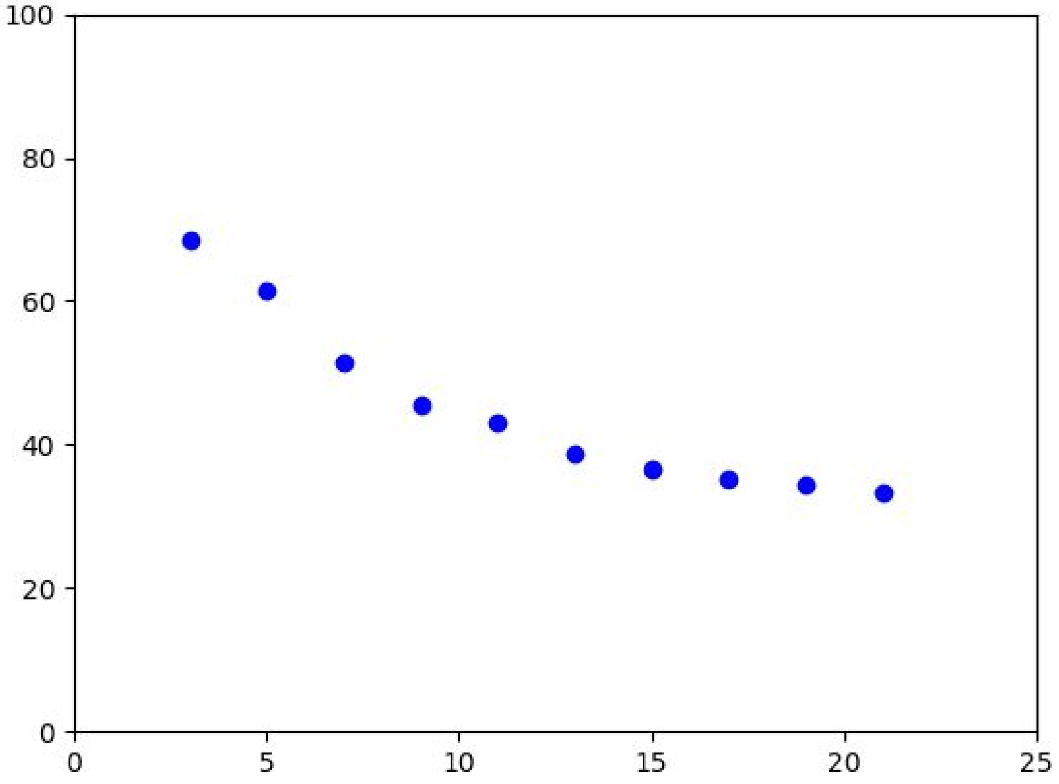
Name : Dhrumil Shah SJSU ID:011428859
Score :0.6859 Rank : 3

Initially, I converted the given input file into csr Matrix and normalized it as I have used Cosine similarity for calculating the distance. For the bisecting K-Means algorithm, I Have maintained list of clusters which is empty initially. Initially there is only one cluster Which contains all the points. Every Time, the bisecting K-Means Clustering Algorithm Calls the K-means Algorithm which runs for K=2 and divide the given cluster into two Clusters. The Sum of Squared Errors is calculated for both the clusters and the one Having the lowest sum of squared errors is added to the list of selected clusters and Then again the other cluster is broken into two clusters and the same algorithm is applied until I receive 7 clusters. There are different methods which I have implemented for finding the clusters and finding the centroids of the clusters. Moreover, for the Dimensionality reduction, I applied truncatedSVD and for calculating the accuracy I am Using calinski herbanze score which gives the ratio between intra-cluster dispersion and inter-cluster dispersion. There is another variable which is number of iterations. For the total number of iterations, it recalculates and recompute the centroid for the cluster.

Here is the graph for the iterations from K=3 to 21.


100
80
60
40
20
0
0
5
10
15
20
25