REPORT

First I created a 2D array which stores all the words present in train file. I created another 2D array which store cluster centres and randomly assigned cluster centres to them using rand function. Using those cluster centres I created another 2D array which stored indexes of the words according to clusters and clusters were created by calculating levenshtein distance of each word with each cluster centre and then store the word in that cluster whose levenshtein distance between word and centre is minimum. After creating and assigning clusters I reassigned updated centroids.

Updation is done by calculating the sum of levenshtein distance of a word with all the other words of cluster and then the word with minimum sum will be called the new centre.

Cluster centres are also called representative strings and K is the number of cluster centres created.

After that I created a Histogram(2-D array dimensions 541 x K) for train data which stores number of times representative strings appear in a particular sms. 541 is the number of sms present in train data. I created a similar histogram(2-D array of dimensions 543 X K) for test data where 543 is the sms in test data.

Taking 1 sms at a time from test data I calculated euclidean distance of1 test sms with all of train data and after that calculate index corresponding to the minimum distance euclidean distance. After that assign the label of train SMS to test SMS. Repeat this for all SMS. The correctly displayed SMS were counted and displayed.