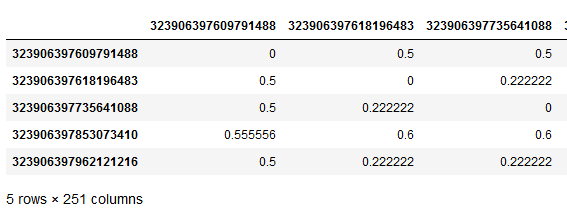
**K-Means**

Dataset

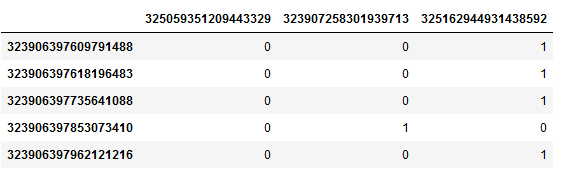
The given tweets.json dataset was parsed to obtain tweet\_id and the corresponding tweet



Jaccard\_distance matrix was calculated for every pair of [tweet\_id1, tweet\_id2]



Clusters were formed for all the seed\_ids by the given matrix where a ‘1’ indicates index\_ids (on the left) belong to the cluster where mean is seed\_id (column)



Note that the tweets do not have the numerical coordinates in Euclidean space, so we computed the "centroid" of a tweet cluster as the tweet having minimum distance to all of the other tweets in a cluster.

The SSE is calculated as follows:

SSE = average(square(jaccard\_dist within a cluster))) for all clusters

Sample Output file:

