Computer Engineering Department

Course Name: CMPE255 – Data Mining

Student Name: Sahitya Mullapudi

Sjsu Id: 011545404

Program 3: Clustering

Semester: Fall,2017

Rank: 11

Accuracy: 0.6562

Dimensionality Reduction: SVD

I have used truncated SVD to reduce the dimensions of the given matrix. But, the accuracy is very low. So, I have not used dimensionality reduction for my final results.

Approach:

1. implemented K means – cosine similarity is used to calculate the distance.

```
Kmeans(csr_mat, clusters)
{
    for j in range (1, 25)
    {
        Centroids = pickCentroids(csr_mat, clusters)
        Clustersmat = assignClusters(csr_mat, centroids)
        recompute centroids(clustermat, csr_mat, cluster)
    }
}
```

2. Implemeted Bisecting Kmeans by using k means

```
Bisecting_Kmeans(csr_mat, kclusters)
{
    Initial cluster = csr_mat
    Cluster_number = 1
```

```
Fill Clist[csr_mat.shape[0]] with 1
For k in range(1, kclusters)
    For j range (1, csr_mat.shape[0])
        If clist[j]==cluster_number
            Target_mat =csr_mat[j,:]
        Else
            Clist[j]=clist[j]+1
    Final_index=Kmeans(target_mat, 2)
    Count no.of indexed per cluster_number
        Target_mat= large_cluster
```

Internal Evaluation Metrics:

Silhouette_score

The Silhouette Coefficient is calculated using the mean intra-cluster distance and the mean nearest-cluster distance for each sample. Library from sklearn is used

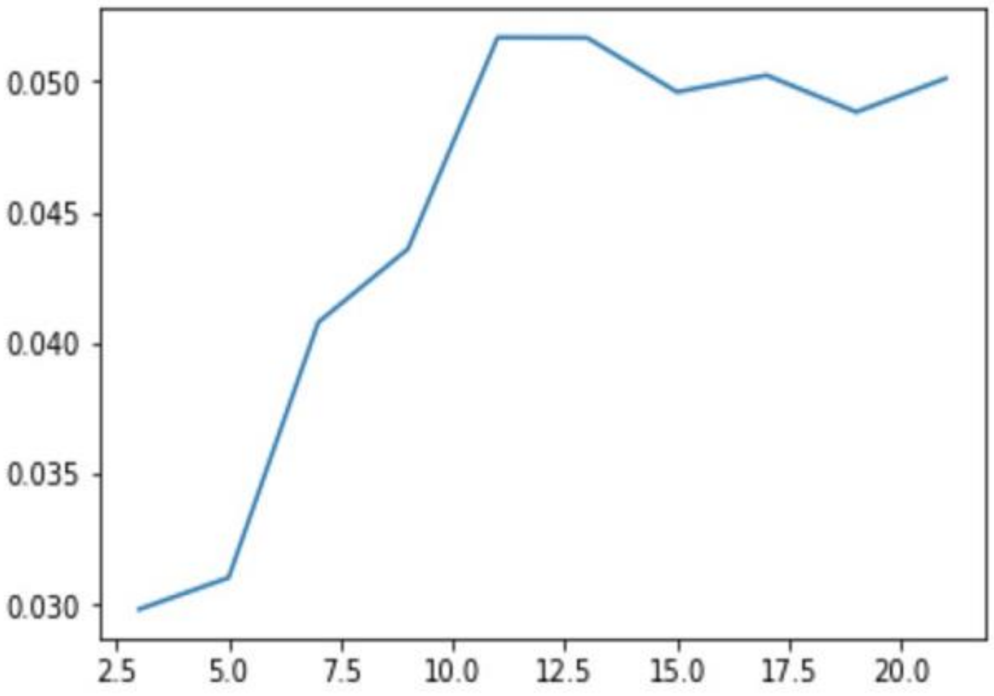
sklearn.metrics.silhouette_score(X, labels, metric=cosine)

score = 0.0238202974302

The low score might be due to using, dimensionality reduction with components = 800, iteration = 7.

Graph

Plotting graph between k clusters and silhouette score with dimensionality reduction – truncated SVD with components=800, iterations=7, random_state=35


0.050
0.045
0.040
0.035
0.030
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0