Report

Programming Homework 1

Adesh Landge

016190486

Objective:

The primary objective is to cluster the text documents by using the bisecting K-means algorithm.

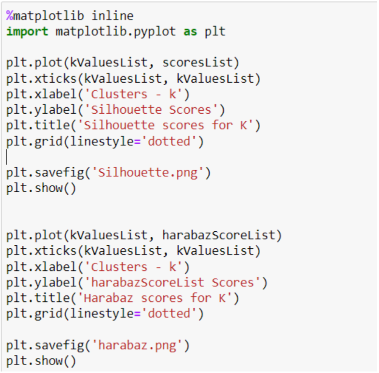
Approach:

1. The given train.dat file contains 8580 text records. This file is read and the given data is converted into a sparse matrix.
2. The compressed sparse matrix is normalized further to its L2-norm form using sklearn's TF/IDF; transforming a count matrix to a normalized tf or tf-idf representation.
3. On top of that, dimensionality reduction is performed on the data matrix using sklearn's TruncatedSVD.
4. In the next step, the goal is to achieve the desired no. of clusters by implementing the bisecting K-means algorithm. The clusters which has the least similarity are returned by applying K-means algorithm. Since, the desired clusters are 7, the final data after creation of 7 clusters is saved to an output prediction file.

Pseudo code:

1. Read the given CSR sparse matrix file.
2. Perform dimensionality reduction on the given data.
3. Compute K-Means algorithm to get two clusters.
4. Compute the similarity of the obtained clusters by performing dot product of the matrices.
5. Store the scores in an array.
6. Find the least score from the array.
7. Perform K-Means on the cluster that has the least score for the similarity.
8. Repeat the steps 4-7 until seven clusters are obtained.

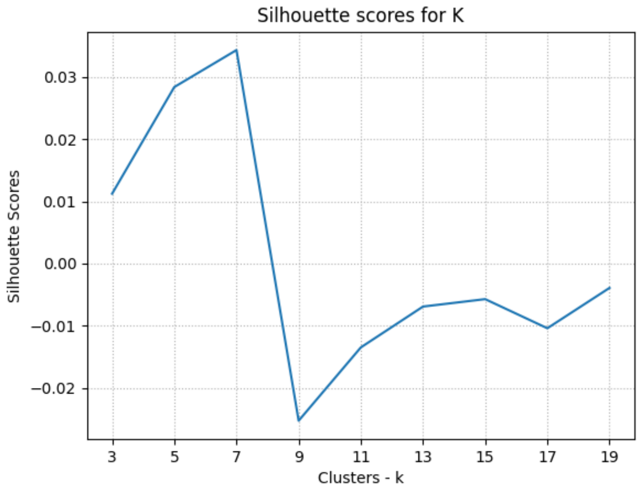
Evaluation metrics:



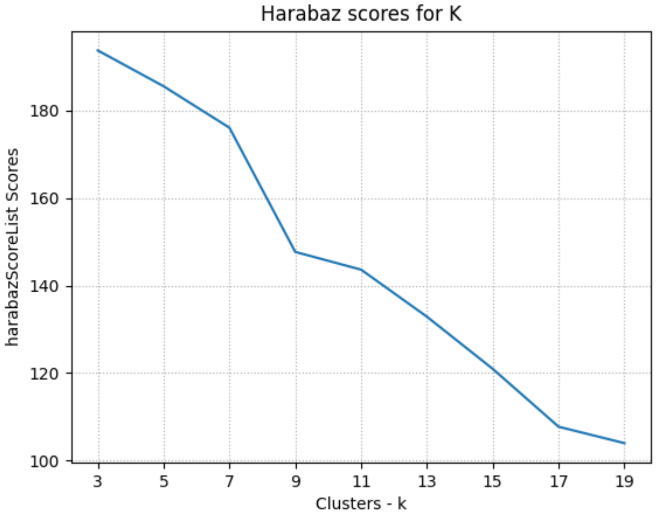
```
%matplotlib inline
import matplotlib.pyplot as plt

plt.plot(kValuesList, scoresList)
plt.xticks(kValuesList, kValuesList)
plt.xlabel('Clusters - k')
plt.ylabel('Silhouette Scores')
plt.title('Silhouette scores for K')
plt.grid(linestyle='dotted')

plt.savefig('Silhouette.png')
plt.show()


plt.plot(kValuesList, harabazScoreList)
plt.xticks(kValuesList, kValuesList)
plt.xlabel('Clusters - k')
plt.ylabel('harabazScoreList Scores')
plt.title('Harabaz scores for K')
plt.grid(linestyle='dotted')

plt.savefig('harabaz.png')
plt.show()
```

1. **Silhouette Score plot:**

   The Silhouette score of 1 means that the clusters are very dense and well separated. Hence, in the below plot for K=7, we have the Silhouette score closest to 1. Therefore, 7 clusters are optimal.



2. **Harabaz Score Plot:**


Harabaz scores for K
harabazScoreList Scores
Clusters - k