Programming Assignment Report

Student: **Vagdevi Junnuri**
Student#: **014926366**

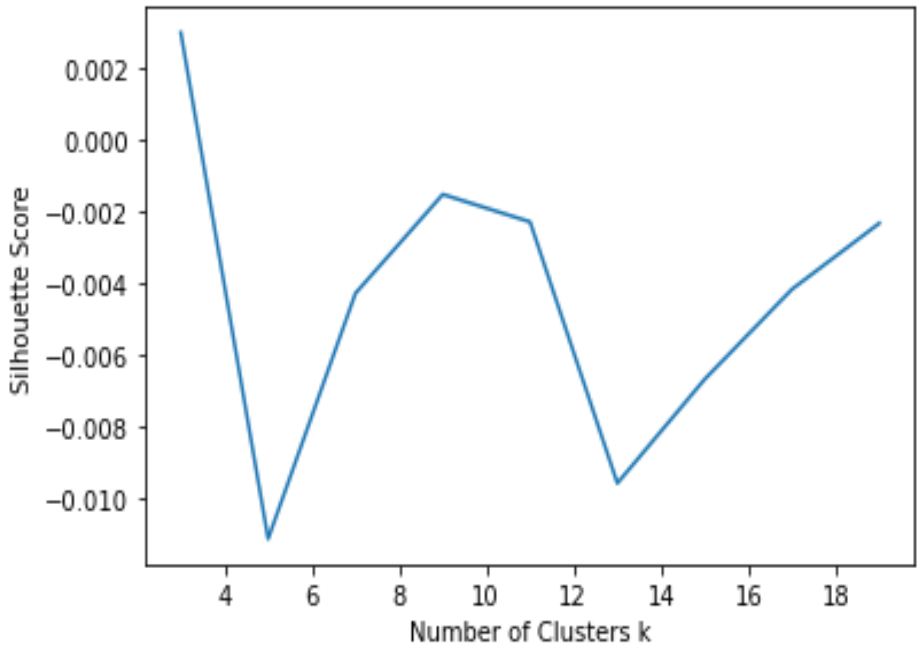
1) Approach :

Pseudocode -

```
        k -> number of clusters
        n_iter -> number of iterations

  Bisecting k-means program start
     Initialize initial_cluster = list()
     Add initial_cluster to clusters_list
     Start while loop(condition : length of clusters < k)
        clusterId_that_should_be_removed = calculate SSE(matrix, clusters_list) and
        find the point with highest SSE value

        clusterId_that_is_removed = clusters_list[clusterId_that_should_be_removed]

        Using k-means the cluster is divided into two clusters Cluster 1 and Cluster 2
        Cluster1, Cluster2 = kmeans(matrix[from clusterId_that_is_removed
        till last index],   n_iter)

        Delete clusterId_that_should_be_removed from clusters_list

        Two clusters returned by k-means is appended to the real(original) cluster.

        Examine index in Cluster1 and
              Append clusterId_that_is_removed[index] to real_cluster1
         Examine index in Cluster2 and
              Append clusterId_that_is_removed[index] to real_cluster2

         Append real_cluster1 and real_cluster2 to clusters_list

(continue bisecting cluster until the k value specified)
```

2) Internal Evaluation metric

As an evaluation metric, Silhouette Coefficient is used.It is used to evaluate the quality of clusters that are created using bisecting k-means.
Calculated Silhouette Coefficient score for number of clusters ranging from k = 3 to 21
Plotted a graph with Silhouette Coefficient on x-axis and number of clusters on y-axis.



Feature reduction -

First the list of documents is converted into a sparse matrix.
In order to scale and normalize the matrix I haved used TF-IDF method.
To scale the sparse matrix IDF is used and normalized the rows of the sparse matrix by their l2-norm.
In order to perform dimensionality reduction, I haved used Principal Component Analysis(PCA).
PCA decomposition is performed on the scaled and normalized matrix.
PCA is used because it can retain most of the information from the original dataset (**variance**) even after the reduction.
It captures the intrinsic variability of data.
The resulted principal components are orthogonal to each other which results in having unique information in each of the principal components.

Sum of Squared Error(SSE) -

SSE is sum of squared differences between each observation and its group's mean.
It is used as a measure of variation within the cluster.
The higher the SSE value, the greater is the variation within the cluster.
SSE is calculated to perform k-means.

By calculating the SSE we know the distance between the points in the cluster and know at which point to divide the cluster.

Bisecting k-means -

Bisecting k-means is a special type of k-means algorithm which is used to find for cluster analysis.It is a type of partitional clustering algorithm which bisects the cluster into two clusters.
In this algorithm, we can start with choosing our own k(number of clusters) value.I have chosen k value as 7.
Before performing cluster analysis, I have preprocessed the data by converting the documents into sparse matrix and did normalization on the sparse matrix.
We first take the whole cluster as one cluster and calculate the SSE value.SSE calculates distance between every point in the cluster and returns the id value of a point which represents the point at which the bisecting should happen.
K-means divides the cluster into two clusters at that point by calculating centroids.
The cluster with the highest SSE value is again calculated for SSE value and performed k-means for further clustering and the process continues till the specified k value.