Assignment 2b Report - Text Mining

1 Introduction

Given a collection of text documents we aim to find similar documents. In order to do that we normalized the text and created a Tf-idf matrix of collection and used cosine similarity to create a similarity matrix. Also applied K-means clustering and hierarchical clustering in order to identify clusters of similar documents.

2 Packages Used - *(Language: Python)*

- **Sklearn**:Package used for constructing Tf-idf, cosine similarity and for K-means
- **NLTK**: Package used for Natural Language Processing.
- **Scipy**: Package which provides function for plotting dendogram and linkage for Hierarchical Clustering.
- **Seaborn**: Used for visualization of data through plots
- **Matplotlib**: Used for plotting of graphs
- **Pandas**: Package which provides Data structure like DataFrame which makes manipulation of datasets easy

3 Dataset

Twenty two text documents were taken all being on the Topic- **The History of web search engines**. Texts are preprocessed and consists of terms corresponding to each document.

4 Methods and Observations

4.1 Tf-idf

Stands for term frequency and inverse document frequency, is a numerical statistic that is intended to reflect how important a word is to a document in a collection. Formula: tf-idf : tf * idf

4.2 K-means

- Used elbow method to find optimal numnber of clusters which came out of to be nine as the maximum drop in SSE error is at point nine as shown in the figure1. Also, this has been verified using Silhoutte coefficient which also attains it's maximum value at iteration corresponding to nine clusters.
- The clusters formed are as mentioned as follows:
  - Cluster 0: *Ass1-1147*
  - Cluster 1: *Ass1-321*, *Ass1-541*, *Ass1-909*

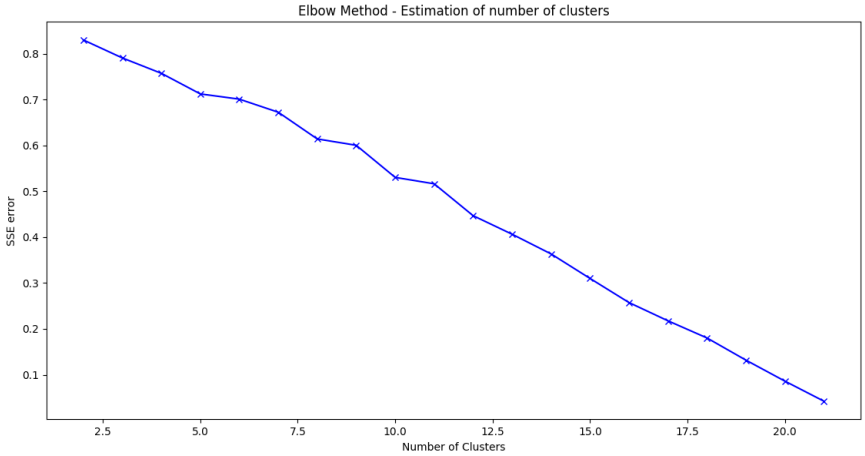

Figure 1: Elbow method

- Cluster 2: *Ass1-1019*, *Ass1-211*,*Ass1-505*, *Ass1-532*
- Cluster 3: *Ass1-734*, *Ass1-1037*, *Ass1-743*
- Cluster 4: *Ass1-1349*, *Ass1-936*, *Ass1-826*
- Cluster 5: *Ass1-1046*, *Ass1-1138*
- Cluster 6: *Ass1-422*, *Ass1-440*
- Cluster 7: *Ass1-808*, *Ass1-202*, *Ass1-817*
- Cluster 8: *Ass1-606*

4.3 Cosine similarity

- It normalizes the lengths of the documents so smaller documents and longer documents have weights of the same order of magnitude.
- According to the similarity matrix as shown in the figure, most similar documents comes out to be *Ass1-422* and *Ass1-440* (Similarity - 0.57) while the most dissimilar documents are *Ass1-202* and *Ass1-1046* (Similarity - 0.025) **Comparison with the previous Similarity matrix using Jaccard coefficient**
  - J.C. does not consider the frequencies of the terms in order to calculate the similarity between documents where as cosine similartity has been calculated using tf-idf vectors of the documents so the similarity between documents has increased as the term frequencies has been taken into account

4.4 Hierarchical Clustering

- **Distance matrix**: It is obtained by calculating *(1-Cosine Similarity)* between each pair of the documents as shown in the figure

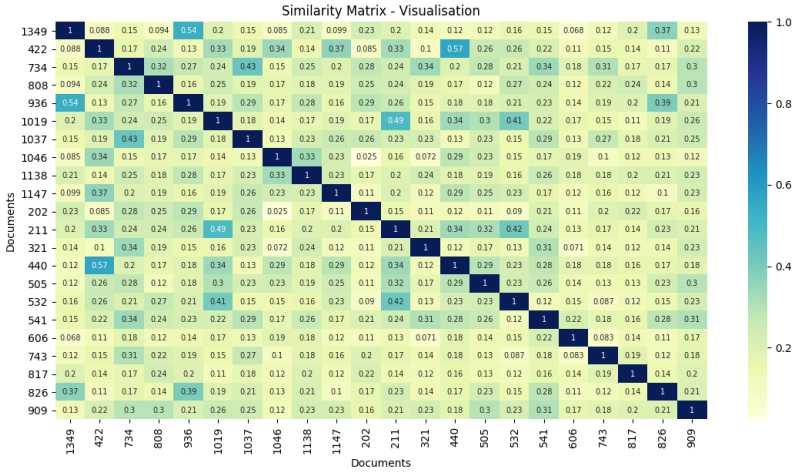

Figure 2: Similarity matrix - Cosine Similarity

- **Linkage Parameter** : Single Linkage
- Dendogram is shown in the figure 3 below where the horizontal axis represents the pairwise dissimilarity between documents

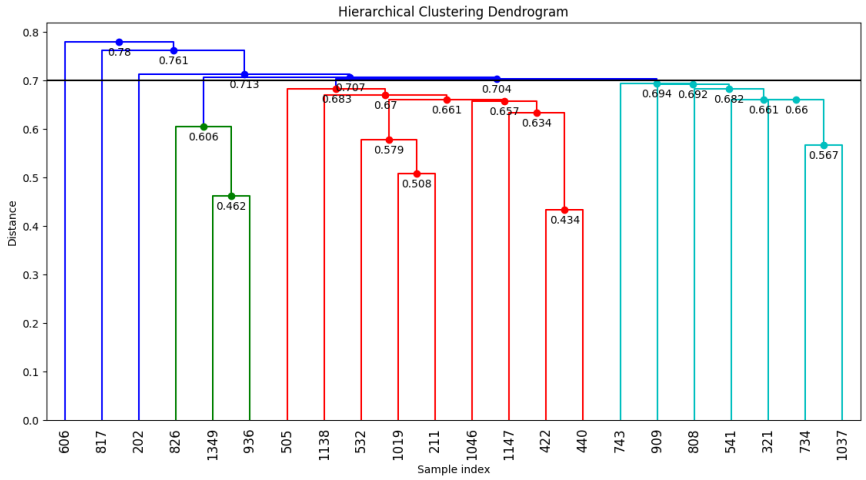

Figure 3: Dendrogram

-

Comparison between K-means clustering and Hierarchical clustering

- K-means tries to minimise SSE by minimising the euclidean distance between the points belonging to the cluster and the center of the cluster.

- In case of hierarchical clustering, cosine distance matrix has been used which is normalised so the different lengths may not neccessarily go into different clusters as in case of k-means clustering
  - For instance, The documents *Ass1-211* (Length - 359) & *Ass1-826* (Length - 240)