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 for each document.

4 Methods and Observations

4.1 Tf-idf

4.2 K-means

4.3 Cosine similarity

4.4 Hierarchical Clustering

- **Distance matrix**: It is obtained by calculating *(1-Cosine Similarity)* between each pair of the documents
- **Linkage Parameter** : Single Linkage
- Dendogram is shown in the figure **??** below where the horizontal axis represents the pairwise dissimilarity between documents

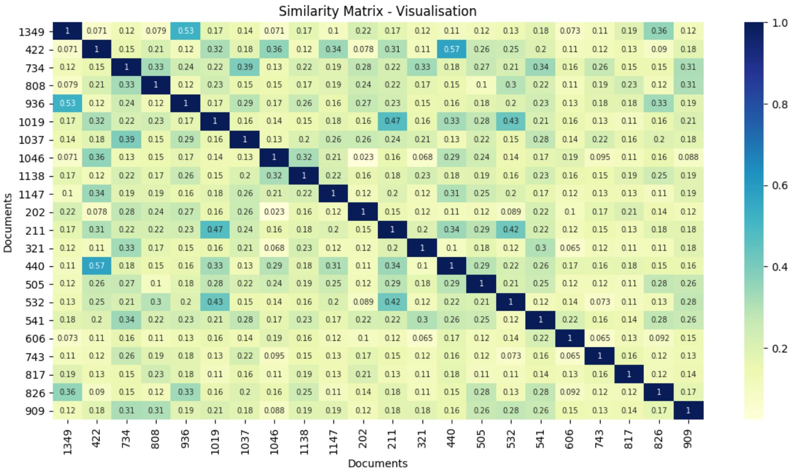

Figure 1: Similarity matrix - Cosine Similarity