Assignment 2c 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 and perform LSA using reduced latent space with 4 dimensions. For each topic we identify the set of 5 top weighted terms. Find the similarity matrix for the documents in the reduced space. Apply hierarchical clustering. Cut the dendrogram at k and identify clusters of similar documents.

2 Packages Used - *(Language: Python)*

- **Sklearn**:Package used for constructing Tf-idf, LSA, cosine similarity.
- **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 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

Figure 1: Similarity matrix - Cosine Similarity

4.3 Hierarchical Clustering

- **Distance matrix**: It is obtained by calculating *(1-Cosine Similarity)* between each pair of the documents as shown in the figure
- **Linkage Parameter** : Single Linkage
- Dendogram is shown in the figure 2 below where the horizontal axis represents the pairwise dissimilarity between documents

Figure 2: Dendrogram