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 Latent semantic analysis(LSA)

We have created Tf-idf matrix and applied LSA to obtain reduced latent space with 4 dimensions. The top 5 words of the four dimensions are as follows:

- *Topic 1*: market, menu, type, aliweb, looksmart
- *Topic 2*: netscap, menu, five, aliweb, instead
- *Topic 3*: purchas, looksmart ,wisenut, dogpil, own
- *Topic 4*: market, netscap, five, year, function

4.2 Cosine similarity

- According to the similarity matrix as shown in the figure, the most dissimilar documents are *Ass1-202* and *Ass1-1046* (Similarity - -0.833) and most similar documents are *Ass1-1037* and *Ass1-541* (Similarity - 0.998)

  Comparison with the previous Similarity matrix using tf-idf:

- LSA considers synonym words as similar so it assigns greater similarity to documents containing similar meaning words whereas in case of Tf-idf, semantics of a word is not considered.

- Using the reduced latent space of four dimensions, it is seen that the similarity between the similar documents has increased and the dissimilarity between the dissimilar documents has also increased as compared to the previous methods used. For instance, the similarity of the most similar documents found using tf-idf (*Ass1-440* and *Ass1-422*) has now increased to 0.9918

- Some documents are found to be almost completely similar (Similarity - 0.998) which indicates that both of them talks about the same topics

- Some documents have negative similarity which indicates that they have terms of different topics

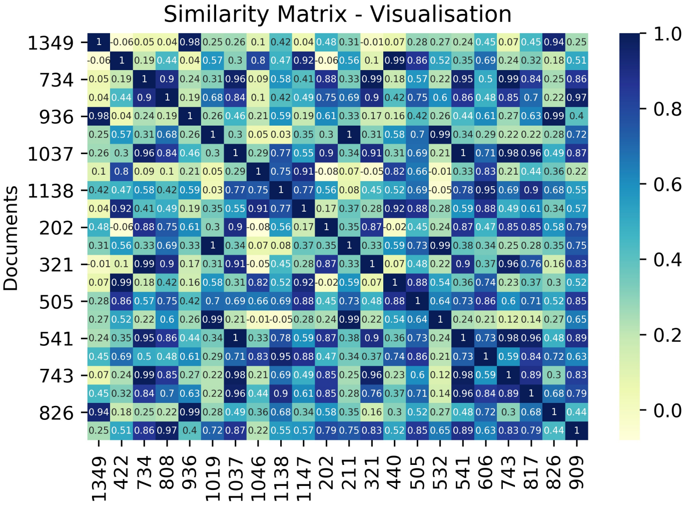

Figure 1: Similarity matrix - Cosine Similarity

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
- The cut indicates that there are a total of three clusters

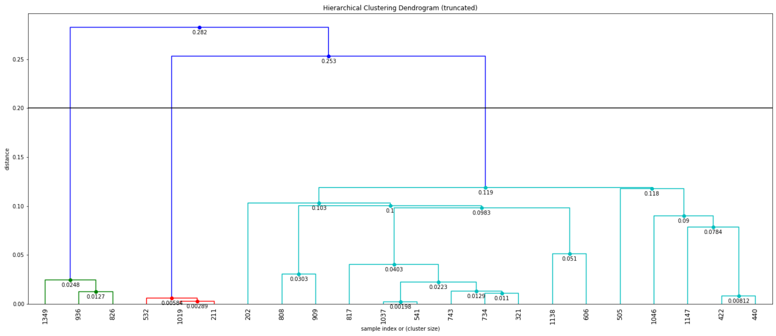

Figure 2: Dendrogram

Comparison with the previous Hierarchical Clustering using tf-idf

- The number of clusters obtained were 6 and in this case, as the similarities between all documents has increased by considering the semantics, the number of clusters has reduced to 3.
- Also, the documents in the clusters are now more closely related where the maximum distance between documents in any cluster is 0.119. Hence, the clusters are tight.