Assignment 2a 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 similarity matrix using Jaccard Index. Also applied hierarchical clustering in order to identify clusters of similar documents.

2 Packages Used - *(Language: Python)*

- **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**.

4 Pre-Processing of Text Documents

1. *Tokenizing* text into sentences and then into words called tokens
2. *Case Folding* - All words were converted to lower case
3. Removing punctuations and digits
4. Removing stopwords - *(Used Standard stop words list for English language)*
    - Stops word list were customized according to the domain by adding *search*, *engine* to the list
5. Performing stemming *(Porter Stemmer)*

5 Methods

5.1 Jaccard Coefficient

- Used to calculate similarity between two sets A and B

$$J.C. = \frac{n(A \cap B)}{n(A \cup B)}$$

- Value varies between 0 and 1 where 0 indicates *no similarity* and 1 indicates *complete similarity*

- **Similarity matrix** is obtained as shown in the figure 1 by calculating jaccard coefficient between each pair of documents

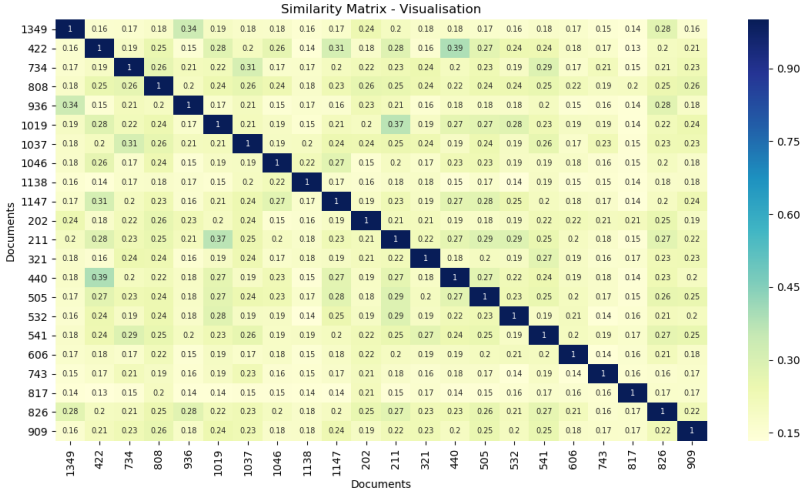

Figure 1: Similarity matrix - Jaccard Cofficient

5.2 Hierarchical Clustering

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

6 Observations

6.1 Jaccard Coefficient

- It does not normalize the lengths of the documents so smaller documents has higher similarity as compared to larger documents in case the cardinality of the set of common words is almost similar.
  - For instance, The documents *Ass1-211* (Length - 359) & *Ass1-826* (Length - 240) has 100 words and 105 words in intersection with the document *Ass1-1349* respectively which is almost same but due to the difference in the length of the documents the first one has 0.198 similarity while the latter one has 0.275 similarity with the document
- According to the similarity matrix, most similar documents comes out to be *Ass1-422* and *Ass1-440* (Similarity - 0.39) while the most dissimilar documents are *Ass1-817* and *Ass1-422* (Similarity - 0.1336)

6.2 Hierarchical Clustering

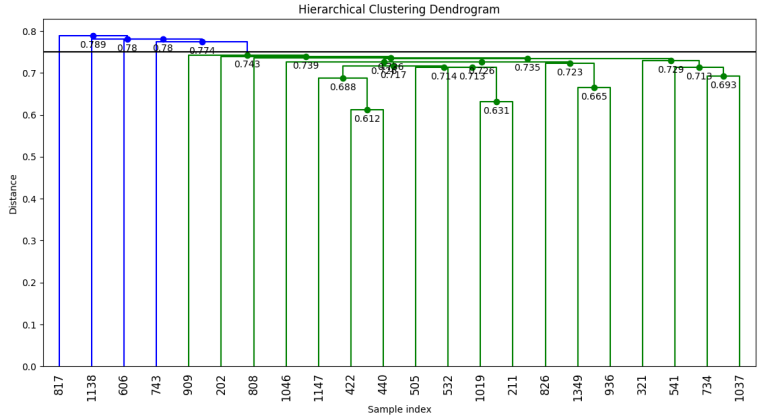

Figure 2: Dendogram - Hierarchical Clustering

- **Elbow Method** (Determining number of clusters)- Maximum acceleration of distance growth in the curve as shown in the figure 3 is obtained from 5 to 6. Hence, Elbow Method suggests that number of clusters would be five. This is also verified with the help of dendrogram where the cut is obtained at the highest jump of distance.
- As observed through the Dendrogram cut at d = 0.75, we have obtained *five clusters* out of the collection of documents
- Documents *Ass1-422* and *Ass1-440* comes out to be most similar (Distance - 0.612)
- Documents *Ass1-817* is most dissimilar to all other documents

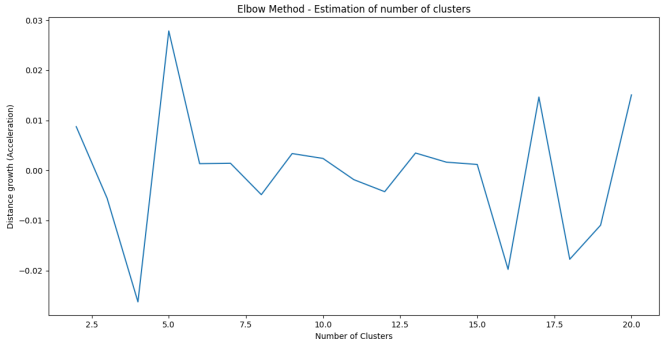

Figure 3: Curve obtained - Elbow method