Text Mining - Assignment 2a Report

1 Package Used (PYTHON) -:

- Package 1

2 Algorithms & Metrics Used -:

- Jaccard Coffecient - Used to calculate the similarity between two sets.
  Formula-:
  $$J.C = \frac{n(A \cap B)}{n(A \cup B)}$$
  Value of Jaccard Coffecient varies from 0 to 1. The value 0 means no similarity and value 1 means completely similar. To find the distance we used the formula $1 - J.C$
- Hierarchical Clustering -
  - Agglomerative Clustering
  - Linkage Parameter used - Ward
  - Distance Metric used - Jaccard Distance
- Porter Stemmer

3 Dataset

Dataset consists of 22 text documents which are write up on History of web search engines. We aim to find plagiarism between the documents.
Pre-processing on data-:

- Tokenization
- Case-Folding
- Removed digits
- Punctuations removed
- Stop words removal - We have removed standard stop words for english and along with it search and engine word is also removed.
- Stemming - We have used Porter stemmer.

4 Observations