Assignment -1

In this assignment, we have explored the fundamental concepts and techniques in the field of Text Retrieval. Text Retrieval is a crucial component of information retrieval systems, and it involves finding relevant documents from a collection based on user queries.

Task Overview

The assignment is divided into several tasks, each building upon the previous one. The main tasks include:

1. Data Acquisition and Preprocessing

   - Obtain raw data from the NASA corpus.

   - Clean the data by removing punctuations, special characters, and digits.

   - Apply tokenization to break the text into words.

2. Stemming and Tag Clouds

   - Perform stemming using the Porter algorithm.

   - Visualize the frequency of words using tag clouds for the 50 most frequent words.

   - Plot the frequency distribution for the 20 most occurring words in the first NASA article.

3. Term Frequency-Inverse Document Frequency (TF-IDF)

   - Calculate term frequency (TF) and TF-IDF weights for each document.

   - Compare the top p stems based on TF or TF-IDF for each document.

   - Compare extracted keywords among all documents and explain differences.

4. Boolean and Vector Models

   - Build Boolean and Vector models based on the top p stems.

   - Provide k queries to each IR system.

   - Compare rankings of relevant articles.

5. Stop Words and Tag Clouds

   - Remove stop words.

   - Perform stemming and create tag clouds for the 50 most frequent words.

6. TF-IDF and Model Comparison

   - Compute TF and TF-IDF for each document.

- Compare the top p stems based on TF or TF-IDF.
- Compare extracted keywords.
- Rebuild Boolean and Vector models.
- Provide the same k queries and compare current rankings.

7. Clustering

- Group similar NASA articles using the TF-IDF matrix.
- Apply a clustering method from scikit-learn.
- Visualize the clustering output.

8. Conclusion

- Finally remarks and conclusions based on the tasks and results.

In order to complete these tasks, I have split up into various python files to implement the Python scripts as mentioned in the AS1 file.

Some of the keywords which was also discussed are

1.Terms (Index Terms)

Explanation: Terms, also known as index terms, are a set of words on which the vector representation of documents is based. In the context of text retrieval, terms represent the building blocks for creating vector representations of documents and queries. Examples of terms include words like "computer," "software," and "fishing." These terms are used to index and retrieve documents in information retrieval systems.

2. Term Weight

Explanation: Term weight is a scalar parameter that represents the significance of a given term within a document. It quantifies how important a specific term is in a document's context. Term weights are crucial for ranking and retrieving documents, as they help determine the relevance of a document to a query. High term weights indicate that a term is essential in a document, while low term weights suggest less importance.

3. Term-Document Matrix

Explanation: A term-document matrix is a matrix that represents the relationship between terms (words) and documents in a corpus. In this matrix, rows correspond to terms, and columns correspond to documents. Each cell in the matrix contains a numerical value representing the weight or frequency of a term in a particular document. Term-document matrices are fundamental for various text retrieval techniques, including TF-IDF and vector models.

4. Document Ranking

Explanation: Document ranking refers to the process of ordering or sorting documents based on their relevance to a particular query. In information retrieval, documents are often ranked to present the most relevant results to users. Techniques like TF-IDF and cosine similarity are used to calculate document scores, which are then used for ranking.

5. Frequency

Explanation: Frequency represents the number of times a word or term appears in a text or document. It is a fundamental concept in text retrieval, as term frequency (TF) is used to measure how often a term occurs in a document. Higher term frequencies often indicate the importance of a term within a document.

6. Corpus

Explanation: A corpus is a collection of documents or texts used for analysis, research, or information retrieval. In your assignment, you mention the "NASA corpus," which refers to a specific collection of documents related to NASA. Corpora are essential for developing and evaluating text retrieval models.

7. Rank of a Word

Explanation: The rank of a word refers to its ordinal position in a list when words are sorted by decreasing frequency. Words that occur more frequently in a corpus have higher ranks, indicating their prevalence and importance. Analyzing word ranks can provide insights into the distribution of terms in a collection.

8. Vocabulary

Explanation: Vocabulary represents the set of all unique words or terms present in a corpus. It includes every distinct term that appears in the collection of documents. Building and managing the vocabulary is a crucial step in text retrieval, as it defines the set of terms available for indexing and querying.

Output Screens

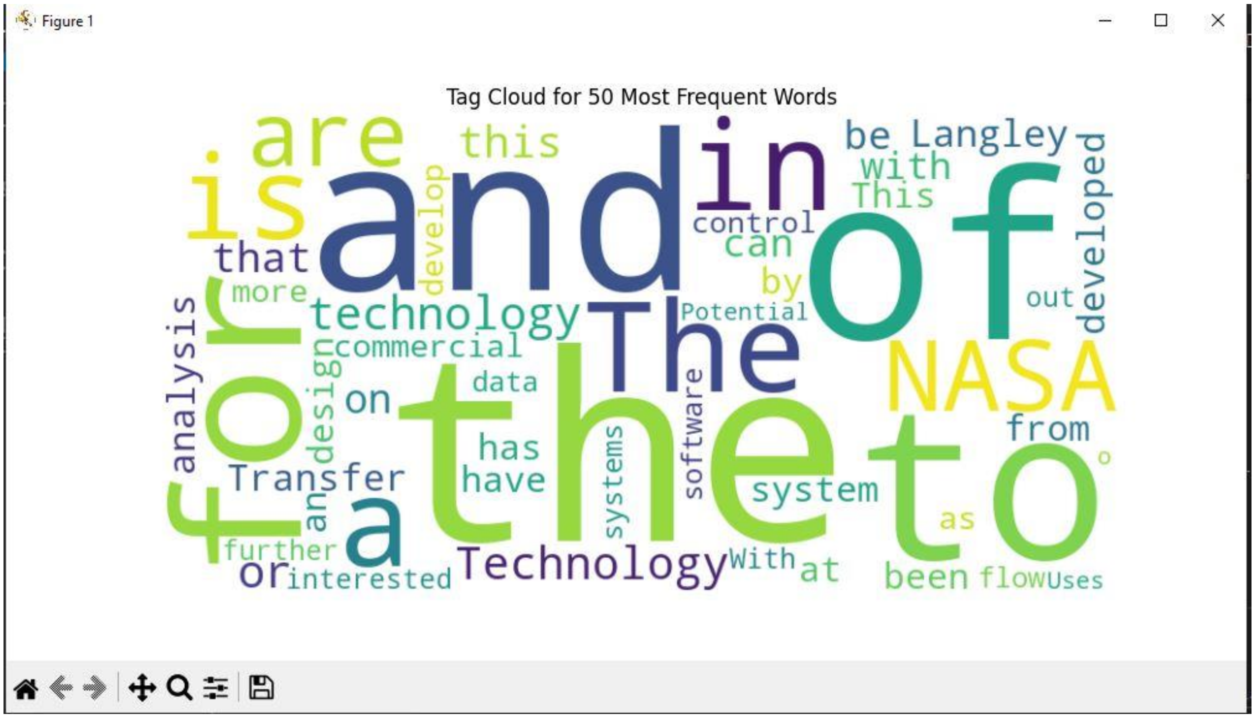


Fig 1: most frequent words before Stemming

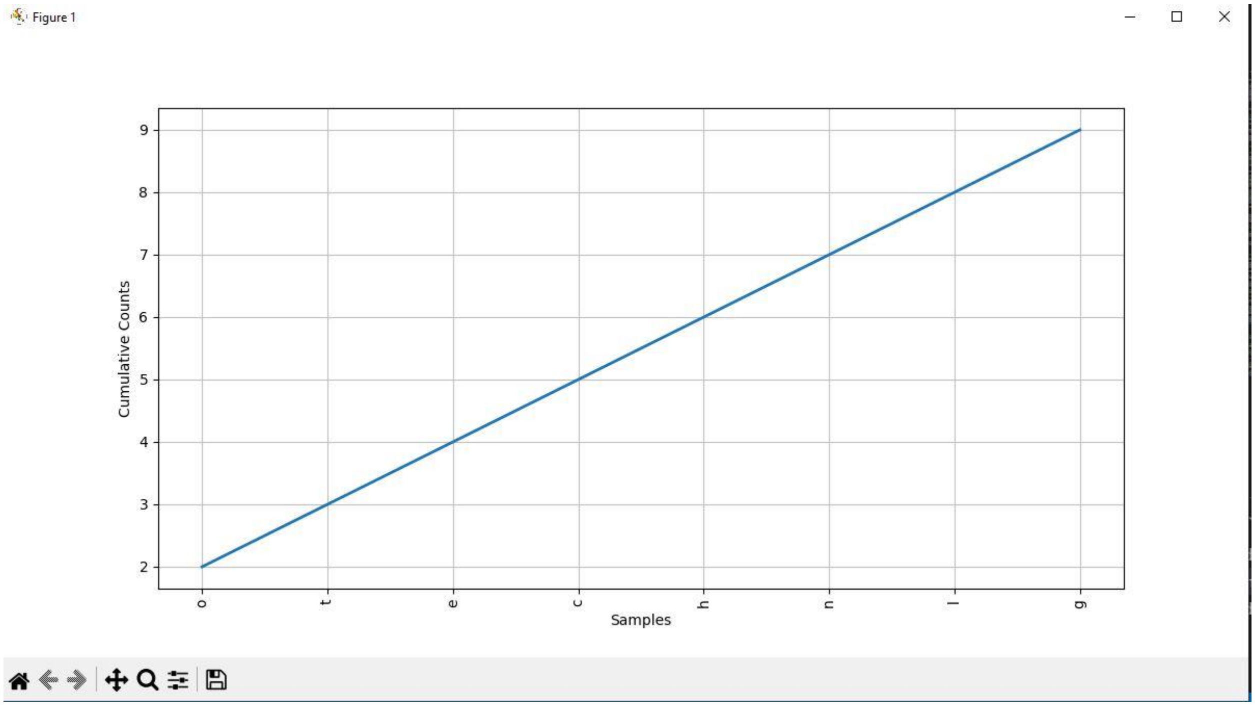
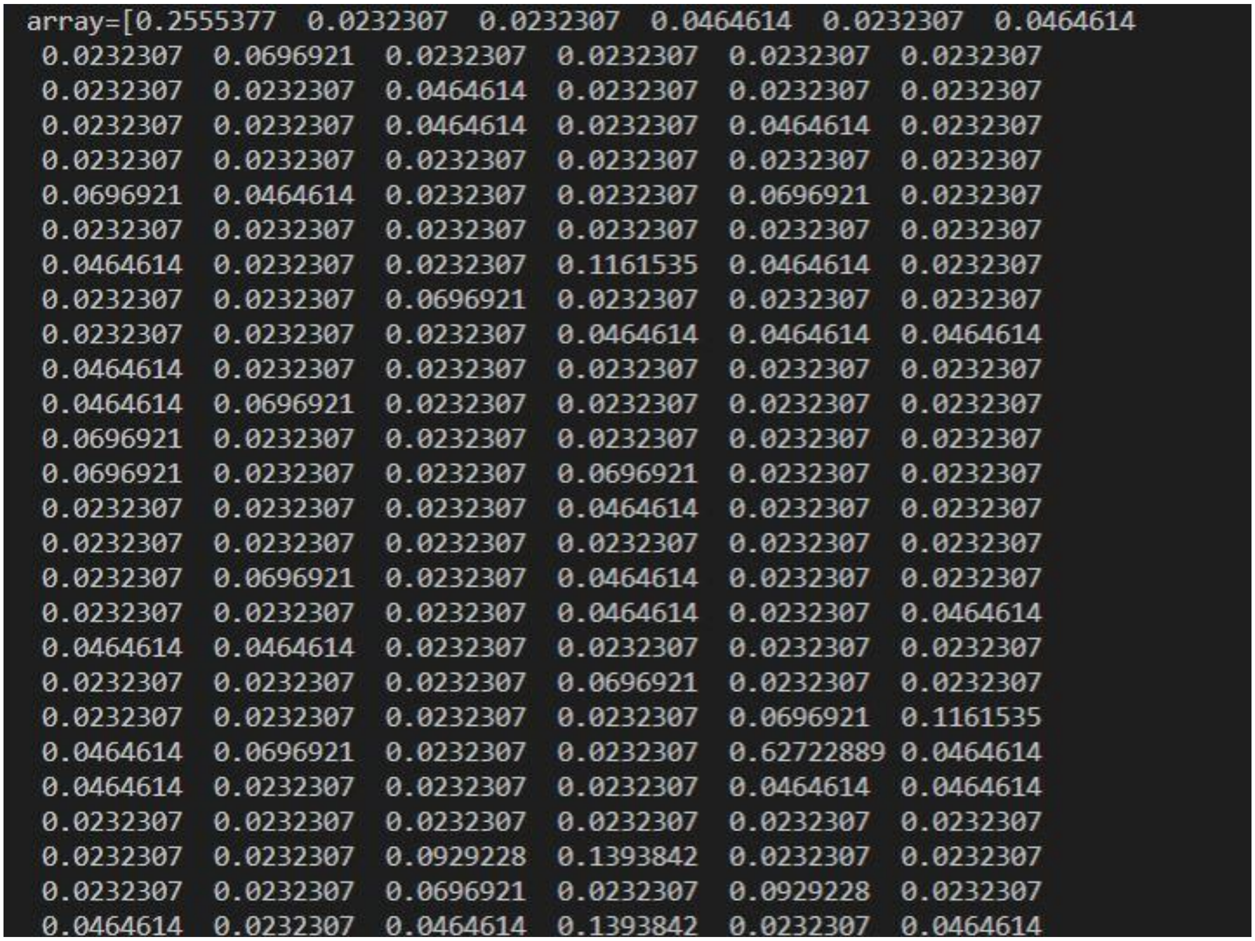


Fig 2: A plot between Cummulative counts and Samples



```
array=[0.2555377  0.0232307  0.0232307  0.0464614  0.0232307  0.0464614
 0.0232307  0.0696921  0.0232307  0.0232307  0.0232307  0.0232307
 0.0232307  0.0232307  0.0464614  0.0232307  0.0232307  0.0232307
 0.0232307  0.0232307  0.0464614  0.0232307  0.0464614  0.0232307
 0.0232307  0.0232307  0.0232307  0.0232307  0.0232307  0.0232307
 0.0696921  0.0464614  0.0232307  0.0232307  0.0696921  0.0232307
 0.0232307  0.0232307  0.0232307  0.0232307  0.0232307  0.0232307
 0.0464614  0.0232307  0.0232307  0.1161535  0.0464614  0.0232307
 0.0232307  0.0232307  0.0696921  0.0232307  0.0232307  0.0232307
 0.0232307  0.0232307  0.0232307  0.0464614  0.0464614  0.0464614
 0.0464614  0.0232307  0.0232307  0.0232307  0.0232307  0.0232307
 0.0464614  0.0696921  0.0232307  0.0232307  0.0232307  0.0232307
 0.0696921  0.0232307  0.0232307  0.0232307  0.0232307  0.0232307
 0.0696921  0.0232307  0.0232307  0.0696921  0.0232307  0.0232307
 0.0232307  0.0232307  0.0232307  0.0464614  0.0232307  0.0232307
 0.0232307  0.0232307  0.0232307  0.0232307  0.0232307  0.0232307
 0.0232307  0.0696921  0.0232307  0.0464614  0.0232307  0.0232307
 0.0232307  0.0232307  0.0232307  0.0464614  0.0232307  0.0464614
 0.0464614  0.0464614  0.0232307  0.0232307  0.0232307  0.0232307
 0.0232307  0.0232307  0.0232307  0.0696921  0.0232307  0.0232307
 0.0232307  0.0232307  0.0232307  0.0232307  0.0696921  0.1161535
 0.0464614  0.0696921  0.0232307  0.0232307  0.62722889 0.0464614
 0.0464614  0.0232307  0.0232307  0.0232307  0.0464614  0.0464614
 0.0232307  0.0232307  0.0232307  0.0232307  0.0232307  0.0232307
 0.0232307  0.0232307  0.0929228  0.1393842  0.0232307  0.0232307
 0.0232307  0.0232307  0.0696921  0.0232307  0.0929228  0.0232307
 0.0464614  0.0232307  0.0464614  0.1393842  0.0232307  0.0464614
```

Fig 3: Array representation of the frequencies

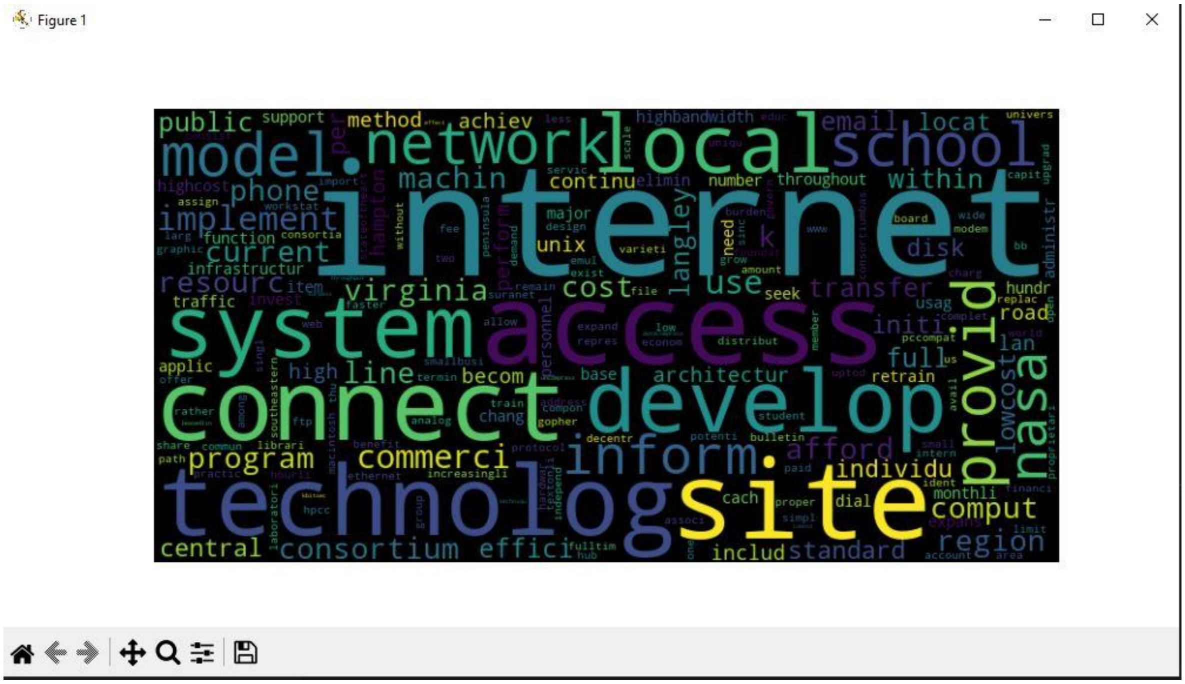


Fig 4: most frequent words after Stemming (unique)

Conclusion

This assignment provides a comprehensive hands-on experience with text retrieval techniques. It covers data preprocessing, vector representations, TF-IDF weighting, Boolean and Vector retrieval models, and clustering. By completing these tasks, we gain valuable insights into how information retrieval systems work and how to evaluate their performance. These key terms and concepts are foundational to understanding and working with text retrieval systems. They play essential roles in tasks such as document indexing, relevance ranking, and information retrieval.