CSE508:Information Retrieval

Sohel Aman Khan (PhD23203)

1 Introduction

This document encapsulates the methodology and insights derived from engaging in Assignment 1 of CSE508. The assignment dissected into three pivotal inquiries, delves into the core facets of Information Retrieval (IR) systems, including Preprocessing, Inverted Index, and Positional Index development and their practical applications.

2 Preprocessing

Preprocessing stands as the cornerstone for refining and preparing textual data for subsequent indexing and retrieval phases. This procedure unfolds through multiple steps:

- **Acquisition of NLTK Datasets:** A prerequisite for fetching stopwords and tokenization utilities.
- **Normalization of Text:** Converting all text to lower case to ensure consistency.
- **Tokenization:** Segmenting text into distinct words or tokens.
- **Elimination of Stopwords:** Discarding frequently occurring words that contribute minimal to IR.
- **Punctuation Removal:** Purging punctuation marks from tokens.
- **Whitespace Cleanup:** Removing superfluous spaces within tokens.
- **Tokens Aggregation:** Merging tokens into cleaned strings for ensuing processes.

Efficiency and precision are paramount throughout the preprocessing stages, aiming for an indexed text of the highest quality.

3 Inverted Index

The inverted index emerges as a fundamental structure in IR, facilitating rapid document retrieval by specific terms. This segment elucidates the creation of an inverted index, encompassing:

- **Document Analysis and Indexing:** Processing documents to map unique terms to their document IDs.
- **Index Storage Solutions:** Adopting a storage solution that optimizes both retrieval speed and space.

4 Positional Index

Enhancing the inverted index, the positional index incorporates the functionality to pinpoint term locations within documents, essential for phrase queries and proximity searches. This assignment section illustrates:

- **Index Expansion:** Augmenting the inverted index to catalog term positions.
- **Query Execution:** Utilizing the positional index to efficiently handle sophisticated search queries.

5 Conclusion

Tackling these assignments underscored the essential components involved in constructing an IR system. From preprocessing textual data to formulating advanced indexing mechanisms, the tasks shed light on the intricacies and strategic considerations vital for efficient document retrieval in IR systems.