Information Retrieval Assignment Report

Abhijay Singh (2021226)

February 9, 2024

1 Introduction

The Information Retrieval assignment for Winter 2024 involved the construction of text processing and indexing systems with a focus on creating an inverted index, handling boolean queries, and establishing a positional index for phrase queries. This report outlines the methodologies and results of the assignment tasks, demonstrating the application of various information retrieval techniques.

2 Data Preprocessing

2.1 Objective

The first task was to preprocess text data to ensure uniformity and remove unnecessary information that could hinder the retrieval process.

2.2 Approach

Preprocessing involved several steps, each critical for simplifying the subsequent indexing and querying processes. The preprocessing script was executed on a dataset comprising 999 documents.

2.3 Methodology

The data preprocessing consisted of:

1. Converting all text to lowercase to normalize case sensitivity issues.
2. Tokenizing the text into individual terms to facilitate indexing.
3. Removing stopwords to omit terms that do not contribute to the search.
4. Stripping punctuation that could introduce noise in the search results.
5. Eliminating any tokens that resulted in blank spaces.

2.4 Results

The text data was effectively cleaned and standardized. The output was a set of documents ready to be used in the construction of the inverted index.

3 Unigram Inverted Index and Boolean Queries

3.1 Overview

The next step was to build an inverted index that would allow for efficient querying of the preprocessed text data.

3.2 Approach

The inverted index was created from scratch without the use of external libraries. The index maps each unique term to the documents in which it appears.

3.3 Implementation

The process of constructing the inverted index and the boolean query mechanism is detailed in the provided code snippets. The index was then serialized using the pickle module for easy retrieval.

3.4 Queries Supported

The system supports a range of Boolean queries:

- AND: Intersection of document sets containing both terms.
- OR: Union of document sets containing at least one of the terms.
- AND NOT: Difference between document sets, excluding terms following NOT.
- OR NOT: Union of document sets containing at least one term and excluding the NOT term.

3.5 Input and Output Formats

The input consists of a series of queries, each defined by a line containing terms and a subsequent line containing operators. The output format presents the query, the count of documents retrieved, and their names.

3.6 Results

The inverted index successfully facilitated the execution of Boolean queries, returning relevant document sets in response to user input.

4 Positional Index and Phrase Queries

4.1 Overview

The assignment's final task was to create a positional index to allow phrase queries within the document set.

4.2 Approach

A positional index was created to map terms not only to documents but also to the positions of the terms within those documents.

4.3 Implementation

The positional index was constructed to support precise phrase queries, allowing users to search for exact sequences of terms in the documents.

4.4 Input and Output Formats

Similar to the inverted index, the input for the positional index consisted of phrase queries. The output detailed the number of documents retrieved and their names.

4.5 Results

The positional index was successfully used to execute phrase queries, enhancing the search functionality of the system.

5 Conclusion

This assignment demonstrated the fundamental processes involved in text preprocessing, indexing, and querying within the context of information retrieval. The implementation of the inverted and positional indices has laid the groundwork for a more comprehensive search system that could be further expanded with advanced features such as ranking algorithms and natural language processing techniques.