CSE508 Winter 2024 Assignment 1 Report

Vishnu Mothukuri - 2021502

Introduction

This report outlines the approaches, methodologies, assumptions, and results of the tasks completed as part of Assignment 1 for the course CSE508 in Winter 2024. The assignment focused on data preprocessing, creating a unigram and positional index, and implementing query processing operations.

1 Problem 1: Data Preprocessing

Approach and Methodology

The first problem required preprocessing a dataset of text files to lowercase the text, tokenize, remove stopwords, punctuations, and blank space tokens. A Python script was developed to automate this process for all files in the dataset. The script utilized the `os`, `string`, and custom functions for text manipulation.

Assumptions

- English language stopwords were considered.
- Punctuation marks were identified using Python's `string.punctuation`.
- Tokenization was simplified to splitting text based on whitespace.

Results

The preprocessing script successfully modified all text files, ensuring a consistent format ready for further analysis. The first five files were printed to demonstrate the preprocessing steps, confirming the script's functionality.

2 Problem 2: Creating a Unigram and Positional Index

Approach and Methodology

For the unigram index, each unique word across the dataset was mapped to the documents it appeared in. The positional index extended this by also recording the word positions within documents, facilitating phrase query processing.

Assumptions

- Words were uniquely identified after preprocessing.
- Document names provided a sufficient index key.

Results

Both indices were successfully created and saved using Python's pickle module for persistent storage. The unigram index supported efficient query processing, while the positional index enabled precise phrase query handling.

3 Problem 3: Query Processing Operations

Approach and Methodology

The script supported operations including AND, OR, AND NOT, and OR NOT for unigram queries and phrase queries for the positional index. Logical operations were implemented through set operations on document lists.

Assumptions

- Queries were preprocessed using the same methodology as the dataset.
- The AND NOT operation was interpreted as excluding documents containing the second term.

Results

The query processing operations were tested with sample queries, demonstrating accurate retrieval of documents based on the specified logical conditions.

Conclusion

The assignment tasks were completed successfully, demonstrating the effectiveness of the developed scripts in data preprocessing, index creation, and query processing. The methodologies applied proved robust for the dataset and query types considered. This report has provided a comprehensive overview of the work undertaken, highlighting the key approaches, assumptions, and results for each problem.