CSE508_Winter2024_A1_2021417_Report.pdf

Overview:

This report presents the solutions and findings for Assignment-1 in CSE508 Information Retrieval, Winter 2024.

Q1. Data Preprocessing:

Approach:

- Developed a Python script (`Q1.py`) for text preprocessing of the provided dataset.
- Utilized the NLTK library for tokenization and stopword removal.

Methodologies:

- Lowercased the text, performed tokenization, removed stopwords, punctuations, and blank space tokens.
- Saved the preprocessed content in a new file for further use.

Results:

- Processed a subset of files to demonstrate the effectiveness of preprocessing.
- Presented original and preprocessed contents for comparison.

Q2. Unigram Inverted Index and Boolean Queries:

Approach:

- Constructed a unigram inverted index from scratch to facilitate Boolean queries.
- Developed a Python script (`Q2.py`) for handling Boolean queries with AND, OR, AND NOT, OR NOT operations.

Methodologies:

- Utilized Python's pickle module to save and load the unigram inverted index.
- Supported generalized queries with dynamic input format and provided accurate results.

Results:

- Executed sample test cases to showcase the correctness of the implemented Boolean queries.
- Demonstrated the input and output format compliance.

Q3. Positional Index and Phrase Queries:

Approach:

- Created a positional index from scratch for efficient handling of phrase queries.
- Developed a Python script (`Q3.py`) for processing phrase queries using the positional index.

Methodologies:

- Implemented text preprocessing and phrase query functionality.
- Employed Python's pickle module for saving and loading the positional index.

Results:

- Executed sample phrase queries to illustrate the accuracy of positional indexing.
- Ensured adherence to the specified input and output format.

Conclusion:

In conclusion, the implemented solutions adhere to the assignment requirements. The provided scripts and functionalities demonstrate effective information retrieval techniques and thorough understanding of the concepts covered in CSE508.
