Information Retrieval course - Mini Project 1 Report

Kimia Esmaili - 610398193

Key Findings:

- The Information Retrieval system successfully preprocesses a collection of text documents, creates an inverted index, and handles both Standard Boolean Queries and Proximity Queries.
- The system provides accurate and relevant results based on the user's queries.

Challenges Faced:

- Preprocessing the documents involved various steps like removing special characters, converting to lowercase, tokenization, and stemming. Ensuring the accuracy and efficiency of each step was a challenge.
- Implementing the Boolean queries required parsing the query input and handling different operators. Ensuring the correct combination of documents based on operators was another challenge.
- Handling proximity queries involved checking the distance between terms in documents, which required careful iteration and comparison.

Enhancements Made:

- The code could be enhanced by implementing additional preprocessing techniques like removing numbers, handling synonyms, or using more advanced stemming algorithms.
- To optimize the inverted index, we implemented Posting Compression. Instead of storing the complete list of document IDs for each term, we store the differences between consecutive document IDs. This reduces the memory footprint of the inverted index, improving efficiency for large collections.
- Additional error handling and input validation could be added to improve the robustness of the system.
- The code could be refactored to use classes and modular functions for better organization and reusability.
- Implementing ranking algorithms like TF-IDF or BM25 could enhance the relevance of search results.
- Overall, the code provides a basic implementation of an Information Retrieval system, but there are several potential enhancements and optimizations that could be made depending on the specific requirements and use cases.