Information Retrieval course - Mini Project 2 Report

Kimia Esmaili - 610398193

Key Findings:

- The code successfully implements basic functionalities like text preprocessing, inverted indexing, spelling correction, and wildcard query handling.
- The system provides improved search accuracy by removing stopwords and suggesting/correcting misspelled queries.
- Wildcard queries allow users to perform flexible searches by accommodating partial or variable terms.

Challenges Faced:

- One challenge was implementing more advanced techniques like Permuterm or K-gram indexing for wildcard query handling. These techniques can further optimize pattern matching and reduce the search space, but they were not included in the provided code.
- Another challenge was handling edge cases and ensuring robustness. The code could be enhanced with additional error handling and input validation to handle unexpected user queries or system errors.

Enhancements Made:

- The code implemented text preprocessing by removing special characters, converting text to lowercase, tokenizing the documents, and removing stopwords. This improves the quality of the inverted index and search results.
- Spelling correction using the Levenshtein distance was added to provide suggestions or automatic correction for misspelled queries, enhancing the user experience.
- Wildcard query handling was implemented to support queries with one or two * symbols, allowing users to perform more flexible searches.

Index Optimization:

- The code doesn't explicitly implement advanced index optimization techniques like Permuterm or K-gram indexing.
- However, the basic inverted index implemented in the code still provides efficient retrieval of documents based on query terms.
- To further optimize the index, techniques like compression, pruning, or caching mechanisms could be implemented to reduce memory usage and improve performance.
- Implementing more advanced indexing techniques can be considered as future enhancements to optimize the system's performance further.

Overall, the code provided demonstrates key findings in terms of search accuracy improvements, challenges faced in implementing advanced techniques, and enhancements made to handle various query scenarios. However, there is room for further optimization and robustness improvements in the code.