Implementing the BSBI Algorithm for Inverted Indexing

About

This project centers on constructing an Information Retrieval (IR) system using the Block-Sorted Based Indexing (BSBI) algorithm. The BSBI algorithm facilitates the creation of an inverted index from a collection of text documents, and the objective is to implement key steps, including document preprocessing, term indexing, and the merging of intermediate index blocks. As an additional focus, students will incorporate gamma coding into the inverted index construction process. Gamma coding, a variable-length coding technique, will optimize the representation of positive integers, improving storage efficiency and speeding up information retrieval, particularly for datasets with skewed distributions where smaller values are more common. This addition provides practical exposure to advanced compression techniques within the context of building an efficient Information Retrieval system.here

Project Overview

Your task is to understand and implement the BSBI algorithm, encompassing the processing of text documents, creating an inverted index, and managing intermediate index blocks. The project is structured to provide hands-on experience in information retrieval and the nuances of building inverted indexes. In particular, during the implementation of the inverted index, it is recommended to utilize gamma coding. Rather than encoding the absolute number of documents, the emphasis should be on gamma encoding the differences between document IDs. Here are the key components and functionalities:

- **Document Preprocessing:**Document Preprocessing: Begin by reading and preprocessing the text documents. This includes converting documents into a word list, tokenization, punctuation removal, and stemming.
- **Inverted Index:** Implement the core BSBI algorithm to build an inverted index. This involves processing documents, indexing terms, and preparing them for later merging.
- **Gamma Coding for Document ID Gaps:** Integrate gamma coding into the inverted index implementation. Use gamma coding to encode the differences or gaps between document IDs.
- **Index Block Merging:** Implement the merging of intermediate index blocks to create a final inverted index. The merging process is critical for handling large document collections efficiently.
- **Implement Bit Manipulation for Gamma Coding(Optional)** For an advanced implementation, consider utilizing bit manipulation techniques to implement gamma coding.

Deliverables

- Functions or methods for handling document preprocessing, building an inverted index, and implementing gamma coding. We would appreciate the addition of comments to enhance understanding for reviewers during the correction process.
- A report summarizing key findings, challenges faced, and insights gained during the project, with a specific focus on the process of implementing the BSBI algorithm.

References

Document and reference the sources, libraries, and tools used in the project. You can refer to the NLTK library for text processing and explore relevant literature on Information Retrieval techniques.