Assignment 3 M1: Index Construction - Developer

Group Members:
Julie Bui (ID: 59422563)
LoLa Alexis Kim (ID: 52727368)
Maxwell Shih (ID: 88195254)

Github link: https://github.com/Skipper321/Index-Search

Analytics:

The total number of indexed documents was 52,961

We implemented checks to skip broken or unprocessable pages, including pages with no content, invalid JSON files, and non-HTML content. To improve runtime efficiency, we also accepted a trade-off by omitting pages larger than 250 KB at the cost of losing some information, as we believed these larger-files would offer limited value to the index.

There were 314,140 unique tokens found.

We decided to implement stemming with the Porter Stemmer, which reduced different variations of the same word to a single root token. Additionally, we applied weights to terms extracted from specific HTML elements such as <title>, <h1>, <h2>, <h3>, <b>, and and added a token cache to speed up repeated stemming operations.

The total size of our index on disk was 136, 551.21 KB

We implemented stemming to consolidate different forms of the same word, reducing redundancy and saving space in memory. Our index processed the documents in batches of 2,000, writing partial indexes to disk and subsequently merging them into a final comprehensive index.

Additional analytics:

	Values
The number of indexed documents	52,961 indexed documents
The number of unique tokens;	314,140 unique tokens total
The total size (in KB) of your index on disk.	136, 551.21 KB on disk
Indexer runtime	~25 minutes
Number of partial indexes created before merge	27 .json files before merging into 1 final .json