Milestone 1 Report: Inverted Index

Jadon Tapp

Overview

This indexer reads HTML pages from the ICS web crawl dataset, parses them using BeautifulSoup, tokenizes and stems all alphanumeric tokens using Porter Stemming, and builds an inverted index mapping each token to its postings. Important words found in titles, headings, and bold text are flagged in each posting.

Index Analytics

Metric	**Value**
Number of indexed documents	1,988
Number of unique tokens	13,122
Total size of index on disk	26,704.10 KB