HomeWork-3 Report

1. **Crawling [Note down steps implemented for each of the below]**

   a. **URL Canonicalization**

   Initially, we are converting the scheme (like HTTP or HTTPS) and the host (the domain name) to lowercase, promoting consistency across varied URLs. Furthermore, default ports for HTTP (port 80) and HTTPS (port 443) are eliminated to prevent redundancy. Relative URLs are transformed into absolute ones if a base URL is provided, aligning them appropriately within the context. The removal of fragments, which typically denote specific sections within a webpage, ensures the focus remains on relevant content during crawling. Moreover, any duplicate slashes in the URL path are eliminated to maintain cleanliness. Additionally, a supplementary normalization step filters out URLs with certain file extensions, such as images or documents, which are generally unnecessary for crawling purposes. This meticulous process streamlines URL formatting, facilitating efficient data extraction from web sources.

   b. **Frontier Management**

   Frontier management in web crawling involves prioritizing URLs using a scoring system. The scoring function evaluates factors such as keyword relevance, trustworthiness of domains, and presence of specific terms like 'data' or 'structure'. URLs are parsed to extract domain and path details, and keywords are stemmed for matching. Trusted domains are identified, and URLs associated with hurricane data are given extra weight based on factors like inbound links. Trusted domain URLs and those with relevant keywords receive higher scores. Finally, scores are adjusted based on the wave number of the crawling process. This approach ensures efficient URL prioritization for effective web crawling. New URLs are added to the queue which are outlinks to the URL currently processed and score is calculated.

   c. **Politeness Policy**

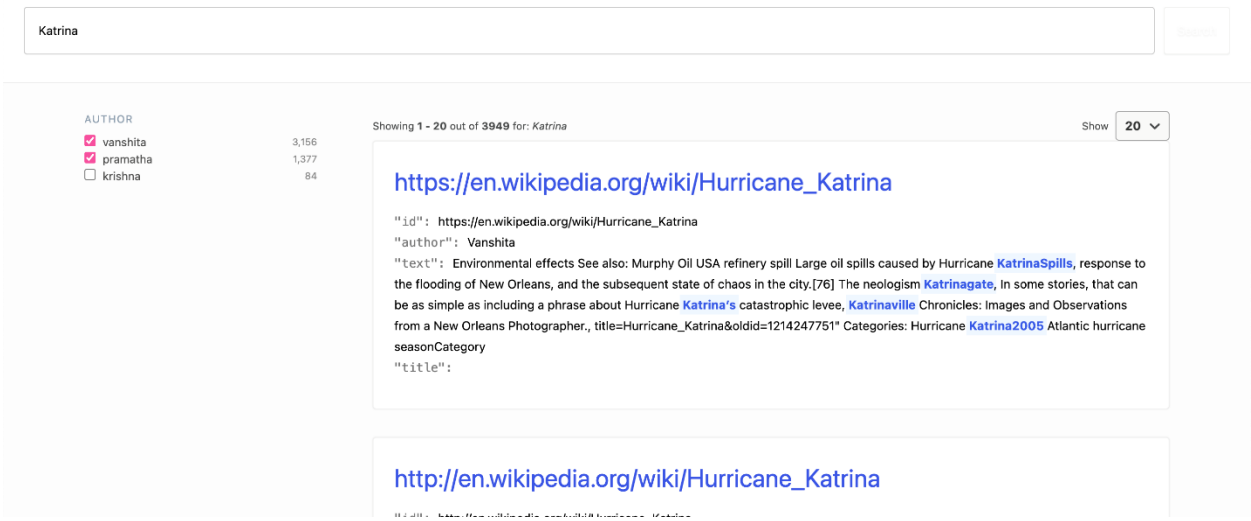
   In implementing a politeness policy for web crawling, a one-second delay between requests is observed to prevent overwhelming servers and respect website policies. Additionally, the crawler adheres to rules specified in the robots.txt file of each domain to avoid accessing disallowed URLs. The crawl process begins with initializing necessary variables and frontier management structures. URLs are fetched from the frontier queue, and if not visited before, they undergo canonicalization and robots.txt validation. If permitted, the page content is fetched and processed, and relevant links are extracted.

   d. **Document Processing**

For document processing, the extraction function parses HTML content using a library, retrieving the title and body text while handling exceptions gracefully. The writing function constructs a document structure and appends it to a corpus file, including the URL, title, and text. Error handling ensures smooth execution, with error messages displayed when necessary. Together, these processes enable the extraction and storage of document information for subsequent analysis. Before indexing, the text undergoes a series of preprocessing steps to ensure its quality and consistency. Initially, all text, including titles and body content, is converted to lowercase to establish uniformity and simplify subsequent processing. Following this, stemming algorithms like the Porter Stemmer are applied to reduce words to their root forms, enhancing the consistency of word representation. Lastly, common stopwords, which contribute little to the semantic meaning of the text, are removed to focus on more relevant content

2. Vertical Search

a. Add a Screenshot of your Vertical Search UI



b. Explain briefly how you implemented it.

The vertical search feature is implemented by integrating the search interface with Elasticsearch (ES) API. Upon receiving the search query from the user interface, the connector fetches relevant search results from the ES index, considering various search parameters such as result fields, facets, and autocomplete suggestions. The search results are then displayed in the search interface, organized with sorting options and facets for refining the search further.

3. Extra Credits Done [Note done what was done for each extra credit]

Search Interface Improvements: The search engine now features an autocomplete functionality, which dynamically suggests search results as users type their queries into the search bar. Furthermore, additional information such as text titles and authors are displayed alongside the URLs in search results, providing users with more context about the content they're browsing. Moreover, query keywords are highlighted within the text snippets displayed in the search results,

enabling users to quickly identify relevant information without having to navigate to individual pages.