ADVANCED COMPUTING CONCEPTS

Project:- **WEB SEARCH ENGINE**

Group ID :- 18

Henil Parikh(110013362)

Jay Pandya(110023841)

Toral patel(110023865)

Richa Kansara(110023778)

FEATURES IMPLEMENTED

1. **HTML TO TEXT :-**

- It will tell your search engine to understand what content is all about and provide your data about page to web search engine.
- Here we are converting HTML files to Text file and storing in local
- Library Used:- JSoup library is used to achieve this task

2. **PATTERN MATCHING :-**

- It can find all occurrence of pattern string in a text string.
- Pattern Matching is the method of testing the expected string seque-nce for the inclusion of certain pattern constituents.
- When exact match if found, pattern matching comes into picture where matched string in links are returned as result
- Algorithm:- For searching the text we have used Brute Force Match.
- Time complexity:- Complexity of Brute Force Match for searching any text is O(mn).

3. **RANKING : -**

- Hash table is used to store file name and sort the data and then display.
- Concept of ranking is nothing but information retrieval and ranking it based on occurrence, which is used in many different real time application and in search engine.
- Most search engine algorithm is used to provide users with accurate and relevant results.
- Linear Data Structure:- HashTable is used to store data where occurrence is stored as value.
- Algorithm:- Here we have used sorting algorithm to implement feature of Ranking in our web search engine.
- Time Complexity:- Time complexity of Sorting is O(n logn).

4. **EDIT DISTANCE:-**

- Edit Distance is a way of quantifying how dissimilar two strings (e.g., words) are to one another by counting the minimum number of operations required to transform one string into the other.
- When exact match if not found, edit distance comes into picture where matched string with minimum distance are returned as result.

- Concept Used:- To extract the link from the website Java Regular Expression is used.

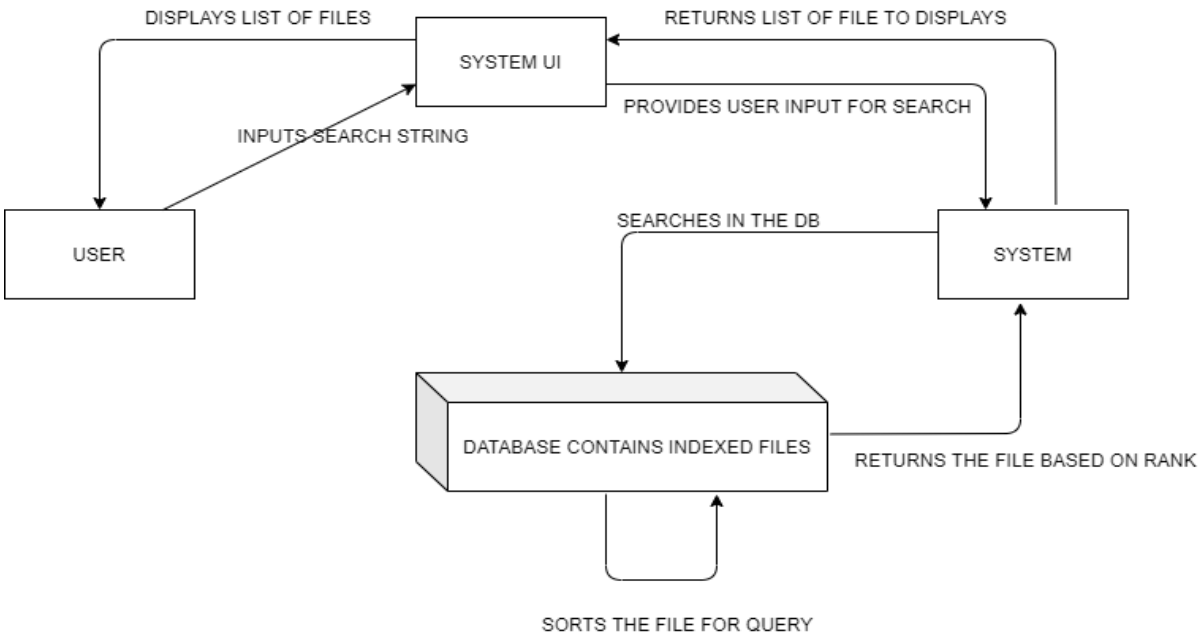
5. **WEB CRAWLER:-**

- It gathers pages from the web and indexes are generated in methodical and automated manner to support queries of search engine.
- Internet is continually evolving and increasing , and there are lots of web pages on the web and it is difficult to find total web pages. Here web crawler concept is used where it has list of known url's.
- They're browsing the webpages at the URLs first. When these webpages crawl, they can discover hyperlinks to other URLs, and add them to the list of pages to crawl next.

6. **SPELL CHECKER:-**

- BFS is used to store unique words and match spelling of particular word that user enters.
- Spell checker is used to check the misspelling of the particular word from the text. This concept is always used in software service like word processor, search engine.
- It scans the text and extracts the words contained in it.
- It then compares each word with a known list of correctly spelled words. This might contain just a list of words, or it might also contain additional information, such as hyphenation points or lexical and grammatical attributes.
- Linear Data Structure:- Binary Search Tree is used to store words as dictionary.

System Diagram:-



References:-

- https://en.wikipedia.org/wiki/Web_search_engine
- Dr Luis Reuda Class notes