Advanced Computing Concepts-COMP8547

Project Report on

Web Search Engine

: SUBMITTED BY :

GROUP 2

1. Mayank Patel (110072211)

2. Urvish Patel (110072574)

3. Rajkumar Patel (110060973)

4. Jay Navnitbhai Patel (110073084)

GUIDED BY:

DR. Mahdi Firoozjaei

Goal: -

The term "Internet Search Engine" refers to a software programme that is fundamentally an Information Retrieval Program that executes a search on a specific database set based on database criteria such as Local Area Network, Wide Area Network, Metropolitan Area Network, or Worldwide Network. They also maintain real time information by running an algorithm on a web crawler. It produces search results in the form of result pages that are properly indexed. Bing, FireFox, Edge and Google are the most well-known search engines on the market. To execute computational computations and locate the suitable search results based on the user's demands, each Search Engine employs a particular algorithm and mathematical formula. Each Search Engine has its own set of features and ways for doing each search in the most efficient way possible, which are listed below.

Functionalities of every search engine: -

- Crawling
- Indexing
- Results
- Storage

Tools and Documents

- Eclipse IDE
- Reference Data from- www.youtube.com
- Jsoup library
- For communication- google meet, Microsoft team and whatsapp

This project is a Search Engine That performs the Search operation and retrieves the results according to the user's needs from the local dictionary. Some of the main concepts and functions used in this project are as below.

1.Web Crawler: - A web crawler is an important and necessary component of every search engine. It utilised to locate website URLs and content, as well as store such information in search engine caches if necessary. It is used to concurrently search a huge number of websites and collect a significant amount of data. It crawls until it has scanned the whole content of the site, including links to internal and external sites.

2.Word Search: - Searching is a method of systematically locating any word or sequence of words in a dataset. In terms of time and space, the different searching algorithms have varying processing speeds. The key distinction between excellent and terrible search engines is the speed and accuracy of their results. In this sense, the search algorithm is extremely important for search engine performance. To find the word from the dataset, we utilised the Boyer Moore method. Because it is considerably faster than using brute force. Because the computing time necessary to do any search with Boyer-Moore is O (nm + s), any search with Boyer-Moore is O (nm + s). So, for pattern matching, we utilised Boyer-Moore in our project.

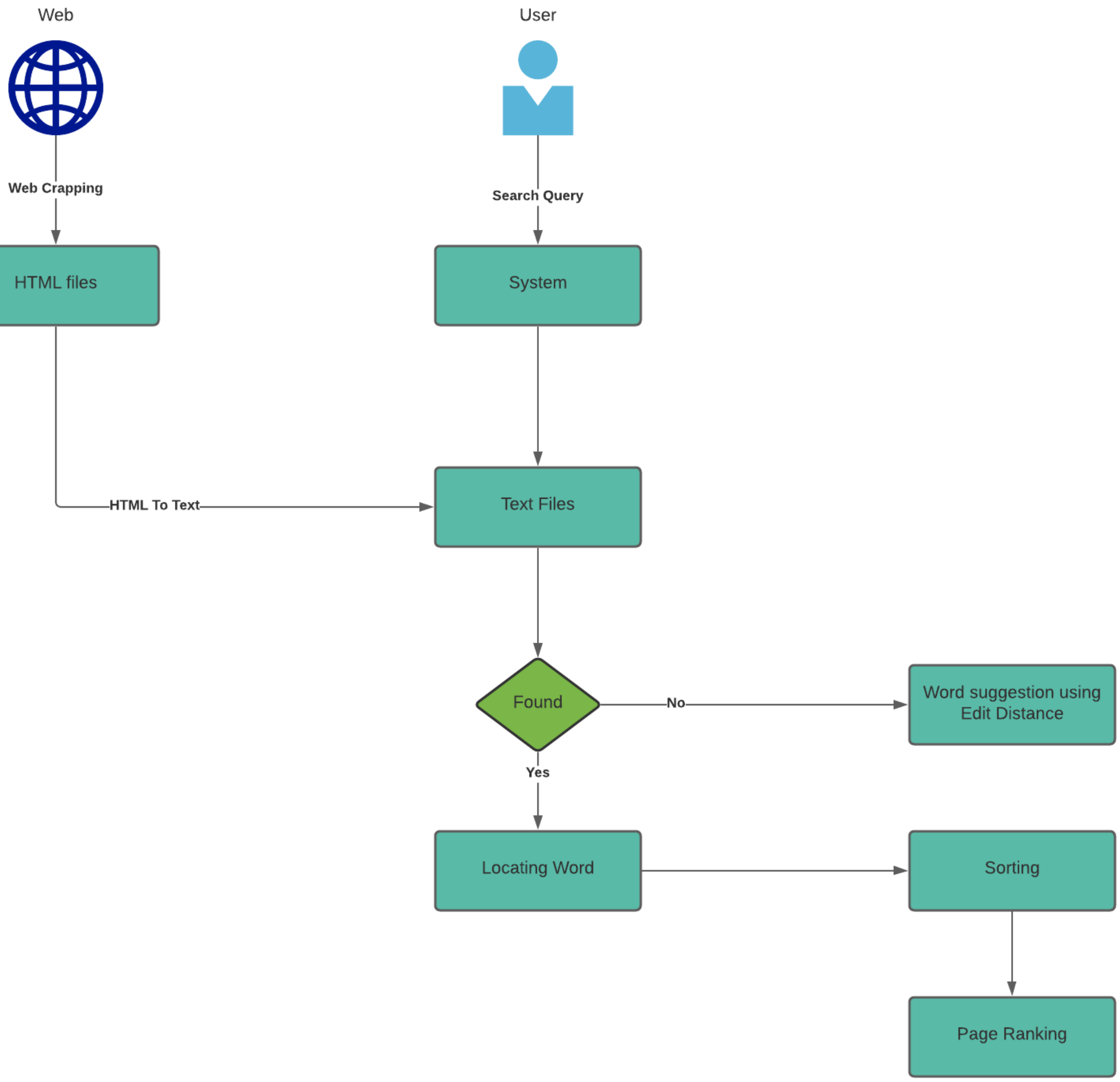
3.HTML to TEXT: - We utilised the Jsoup package to convert HTML pages to text files, and then we used those text files as a dataset for our search engine, on which we performed the searches.

4.Ranking (Indexing): - After the crawler has browsed the dataset's contents, the material is indexed based on the occurrences of keyword phrases. It organises the material so that it may be accessed quickly and easily. The material was saved in a hash table, and the results were sorted using the java.collection.sort inherent sorting algorithm.

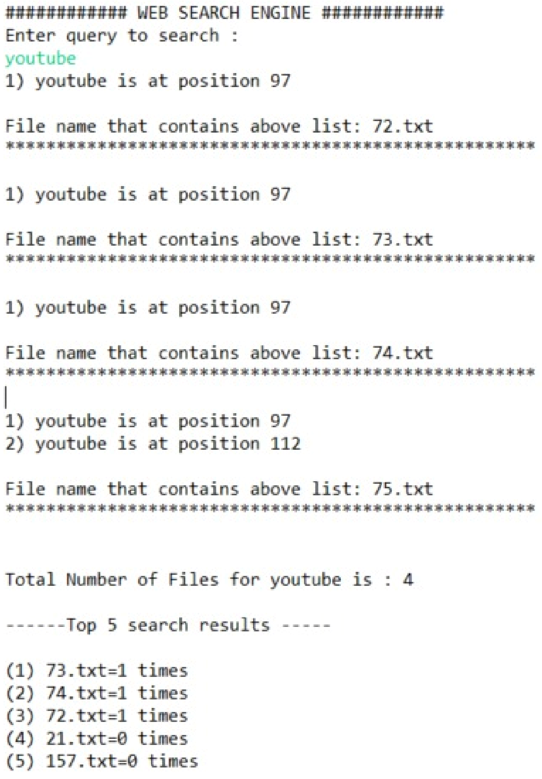
5.Word Suggestion: - In the search engine market, the term "recommendation" is always the most commonly utilised idea. Another metric of your search engine's effectiveness is how well it recognises the user's mistake and provides the appropriate result despite the user's error. In this project, we utilised the Edit Distance technique to make this capability work. It reads the text and retrieves every single word. Then it compares those words to our dataset's correctly spelt words and makes the right suggestion.

So, these are all of the project's key functions. The search engine uses a web crawler to discover the term input by the user, then performs searching algorithms, indexing, and ultimately displays the results for the user's requested search. If the term isn't discovered, it returns a word suggestion as a replacement.

System Flow:


Web
User
Web Crapping
Search Query
HTML files
System
HTML To Text
Text Files
Found
No
Word suggestion using Edit Distance
Yes
Locating Word
Sorting
Page Ranking

Output:



```
############ WEB SEARCH ENGINE ############
Enter query to search :
youtube
1) youtube is at position 97

File name that contains above list: 72.txt
****************************************************

1) youtube is at position 97

File name that contains above list: 73.txt
****************************************************

1) youtube is at position 97

File name that contains above list: 74.txt
****************************************************

1) youtube is at position 97
2) youtube is at position 112

File name that contains above list: 75.txt
****************************************************


Total Number of Files for youtube is : 4

------Top 5 search results -----

(1) 73.txt=1 times
(2) 74.txt=1 times
(3) 72.txt=1 times
(4) 21.txt=0 times
(5) 157.txt=0 times
```

References: -

1) Class Slides of Advance computing concepts.

2) "Boyer Moore Algorithm for Pattern Searching - GeeksforGeeks", *GeeksforGeeks*, 2020. [Online]. Available: https://www.geeksforgeeks.org/boyer-moore-algorithm-for-pattern-searching/ [Accessed: 03- Apr- 2020].

3) Algorithms and Me. (2019). The minimum edit distance between two strings | Algorithms and Me. [online] Available at: https://algorithmsandme.com/minimum-edit-distance-between-two-strings/ [Accessed 22 Nov. 2019].