Project Report

Bangla Search Engine
“ONNESHA”

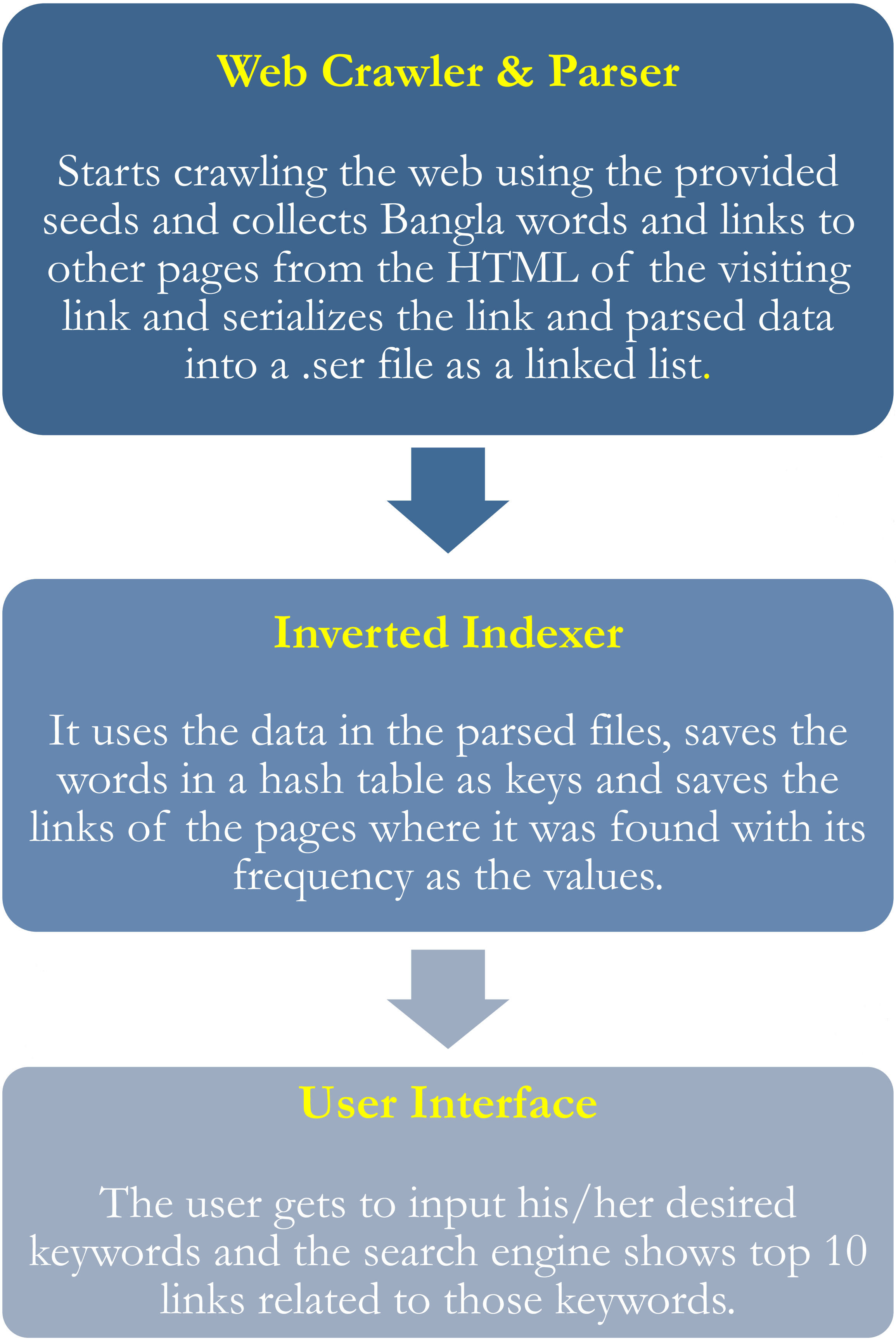
Submitted By -
Khandker Aftarul Islam
Student ID : 1605063
Sk. Md. Mostofa Mohiuddin Sunan
Student ID : 1605073

Supervised By –
Mohammad Saifur Rahman
Assistant Professor
CSE, BUET

Objectives :

- ❑ Developing a fully functional and effective Bangla search engine
- ❑ Learning the basics of a search engine
- ❑ Learning advanced features of Java SE
- ❑ Learning how to implement "Multi-threading" and "Networking" in a project
- ❑ Finding out the problems regarding queries in Bangla and other problems of using UNICODE and their solutions
- ❑ Understanding real world IT-based problems and how to find their solutions

Project Description :



Web Crawler & Parser

Starts crawling the web using the provided seeds and collects Bangla words and links to other pages from the HTML of the visiting link and serializes the link and parsed data into a .ser file as a linked list.

Inverted Indexer

It uses the data in the parsed files, saves the words in a hash table as keys and saves the links of the pages where it was found with its frequency as the values.

User Interface

The user gets to input his/her desired keywords and the search engine shows top 10 links related to those keywords.

Accomplished Tasks :

- We have developed a customized parser which is very efficient at collecting all the Bangla words and links from the HTML file of the visiting link.

- For getting optimum result we have created our very own simple stemmer which removes the “bivokti’ from the collected word and returns the stem of that word.

- We have collected a list of stop-words from Github’s awesome Bangla toolkits; exclusion of these words makes the searches more optimized.

- We have serialized the collected words as linked lists into a .ser file which saves memory and saves time as well because by following this process we can get rid of some time-consuming string operations during indexing.

- Last but not the least, according to Saifur sir, we can claim that our search engine works better than “pipilika” search engine which means it really produces meaningful results.

Proposed Improvements :

- One of the major drawbacks of this project is consumption of massive amount of RAM. The dictionary, which contains all the words and links, is a very big object and its size increases as we index more files and we have to keep it in the RAM on runtime for doing the search. We can solve this problem by implementing database or we can also use disk memory or by increasing RAM.

- Another major drawback is, when the user enters random keywords containing more than one word, sometimes the search engine cannot provide relevant result because the results are provided based on the weight which is determined by the frequency of the words in a link whereas the rarity of a word should also be considered for producing more relevant search results.

- Our stemmer is really a very simple one. It can be improved a lot. But this improvement will require good knowledge of Bangla grammar. The list of "Bivoktis" can also be updated.

- The user interface is also a very basic one. That's why, adding a GUI will give it a nice look.

Instructions :

- Use IDEA Intellij to open and run the project.

- Make sure that there is a folder named “assets” inside the project’s folder and a folder named “ParsedFiles” inside that folder.

- Make sure that the assets folder should contains
  - BanglaStopWords.txt
  - Bivokti.txt
  - ParsedFileNumber.txt
  - indexedFileNumber.txt
  - LinksToVisit.ser
  - LinksVisited.ser
  - Dictionary.ser files.

- If any of the above mentioned files or folders is missing check the following link this link for those files. https://github.com/KAIMonmoy/BanglaSearchEngine_Onnesha_

- This project consumes a huge amount of RAM. So please make sure that your computer has at least 2GB RAM.