SEARCH ENGINE

Amit Vadnere(10442085)

OVERVIEW

The program has a hardcoded base URL which it scraps to get the hyperlinks associated with that URL'S. The scrapped URL's and the base URL is scrapped one by one and stored into tries. The tries data structure is used to find the word and the associated URLs.The page ranking is based on the sum of the number of occurrence of individual words in a URL.

APPROACH

The initial base URL is scrapped using the JSOUP library of java which basically extract all the hyperlinks associated with the base URL. The extracted hyperlinks are stored in ArrayList with the base URL. This ArrayList is traversed one by one. At each index of the ArrayList, the URL's are scraped and the scraped content is filtered for stop words and special symbols. The program filters out all special symbols so as to store only alphanumeric characters in the scraped content. Moreover, the scrapped text is converted to lowercase. After filtering the text is split into words and added into tries. The tries are made up of nodes which consist of an attribute called temp Object which basically stores the 2 attributes i.e.isWord and the map. The value of attribute isWord is true for the node corresponding to the last character of the word in the tries and the map is basically hash map which stores the link corresponding to the word as a key and number of occurrence of the word as a value. There is also a getIndex method which gets a unique integer value for each character or number. After the addition of the words into tries from each hyperlink, the program asks for the user the input query to search. The program searches for all the words of the sentence one by one and gets the frequency count associated with the word. This frequency associated with each word is stored into a hash map with link as key and frequency as value. For each word match in the tries, the frequency is updated corresponding to the link. **The program then sorts the hash map and print the value of key in ascending order of the frequency/value on the console. So the page URL with 1st rank will lie at the bottom.**

BOUNDARY CONDITIONS

The program accepts alphanumeric text and searches for the whole word rather than prefix or suffix.

DATA STRUCTURE

The program uses **tries** as the major data structure for storing and searching of words and **hash Map** to support page ranking and multiword search.

HOW TO RUN?

To execute the program you will need the JSOUP jar file and then the program can be executed using basic JAVA commands that are javac and java for compiling and executing the command.

The program will take initially some time in scrapping the documents this due to a large number of hyperlinks in the base URL. The scraped percentage can be viewed on the console.