Search Engine:

The World Wide Web contains a huge collection of text documents (web pages). Information about these pages is gathered by a program called a web crawler, which then stores this information in a special dictionary database. A web search engine allows users to retrieve relevant information from this database, thereby identifying relevant pages on the web containing given keywords.

Technology: Java - Eclipse Editor – Dynamic Web Project
External Archive: Jsoup-1.10.1.jar

Described below how Search Engine Project works step by step:

Step 1: Run project on the Server (Ex. Apache Tomcat Web Server). That will redirect to index.html page. Click on Add new URL link on the page, that redirects to index2.html page. On this page add URLs of web pages for crawling and Click on Add button that redirects to WebCrawler Java Servlet.

Step 2: WebCrawler Java Servlet passes the entered URL to InformationCollection java class. Then, InformationCollection class passes one by one up to 5 URLs including URL that we have entered and another 4 URLs of hyperlinks that are already in entered URL web page to ResponseHandle java class for handling HTTP requests of webpages of those URLs.

Step 3: In ResponseHandle class using jsoup.jar library, that will handle all the requesting and handling of all http requests and scrape and parse the HTML of the URL passed by InformationCollection class to the String. Then, ResponseHandle class pass the String of the web page to URLProcess java class for storing all the words of the page in well manner.

Step4: In URLProcess class, make compressed trie so that it can arranged words of the web page in well manner and filtering out stopping words such as symbols, single alphabet, pronouns, articles, prepositions and HTML tags.

Step 5: Step 3 and 4 repeats for all 5 URLs and stored all the words of web page in compressed trie including URLs. Now go to main page of the project index.html to search URL using keywords. Enter any keyword and click on Search button that redirects to WebSearch java Servlet.

Step 6: WebSearch Java Servlet passes the entered keyword to ResponseHandle java class using search function that will checked for all pages that are already visited or not and search the keyword in compressed trie and return all the URLs containing that keyword to WebSearch Servlet. Then servlet will print out all the URLs as in link format so that by clicking on them we can redirect to that page.