**PROJECT TITLE :-** MY SEARCH ENGINE (TOGGLE)

**DESCRIPTION :-**

In these project I have to make a search a text with precision and accuracy speed. I need to search keyword and phrases from a given data set.

Apache TIKA is a toolkit that detects and extracts the metadata and text over a thousand different files. All of these files types can be parsed through a single interface making it useful for search engine indexing. In my project I used latest version of Apache TIKA 1.16 .

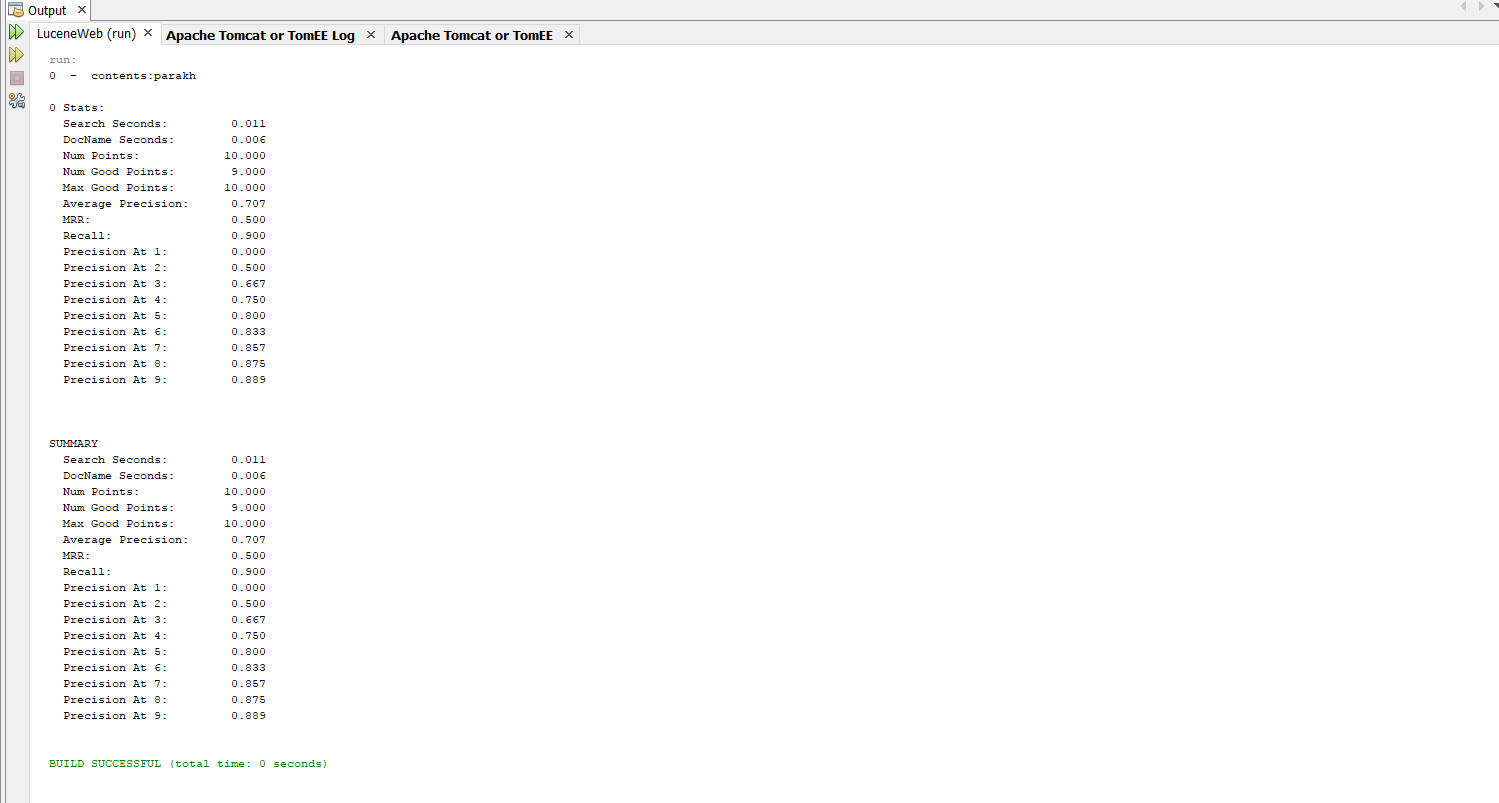
Apache Lucene is a high-performance, full-featured text search engine library written entirely in Java. It is a technology suitable for nearly any application that requires full-text search, especially cross-platform. Lucene offers powerful features through a simple API like scalable and high-performance indexing; powerful, accurate and efficient search algorithms and cross platform solution.

Here I first parsed all the document in metadata using Apache TIKA and than extract the data from that document for indexing and searching using Apache LUCENE data structure. Here I used different properties of lucene for indexing at the world level and search the keywords from that indexed documents. Here I also implemented a intuitive UI to facilitate the search using Servlet and JSP. Here I used golden set of documents in my dataset like pdf, docx, xml, html, txt, json. I also implemented the option for searching multiple queries at one time using conjunction, disjunction and negation operator. The search page will display the name of the file, path of the file, snippet and the ranking score of that document.

**ANALYSIS:-**

For analysis I calculate the precision and recall of all the indexed document using TrecJudge. Precision is the fraction of relevant instances among the retrieved instances, while recall is the fraction of relevant instances that have been retrieved over the total amount of relevant instances. Both precision and recall are therefore based on understanding and measure of relevance. TrecJudge compares the indexed document and calculate the precision and recall for a particular search query.

The average precision for all the documents is 0.707 and recall comes out to be 0.900.



**STEPS TO RUN THE CODE:-**

1. First run the code.
2. Then it will call the index method for indexing all the document.
3. Enter the word or phrase you want to search.
4. Then it will search for the query by calling searching method and it will return an list of all the documents.
5. Then it iterate over the list and display the result for that query.
6. If you want to search for more repeat the speed again.