Report

Programming Language used :- Java , JSP , Python

Architecture :-

This assignment has a web UI where user inputs query and gets ranked search results along with search time .

Extra Features Added :-

1) The server side of the program has a LRU cache implementation which can store upto 5 last used query and response designed to give a quick response to user .

2) The search engine has an inbuilt spelling checker which can predict the right text user might have entered to search .

Implementation Details

At the start of web server , all the text documents which are inside a folder **text** are picked , cleaned , tokenised and stemmed . Using all the tokens we create an inverted index whose postings store document id and term frequency of the token in the document . A map between document id and document name is also stored for future reference .

Once the index structure is created and user enters a query , at first we check if there is any error in spelling of query .This is done using python script which has an edit distance implementation and compares the spelling with the text in big .txt (Reference :- Peter Norvig's implementation).

If the query is correct , the query is cleaned , stemmed and stop words are removed .Now we check if the query is present in our LRU cache . If we get a hit a direct response is sent to user , else we need to do further processing . Next , we create a normalised tf-idf representation of query .We also create a union of all documents in which even single word of query is present . We create a normalised tf score of these documents .Multiply and add both scores above for each document and sort them in descending order and print the result .

Start time for search and end time for search is maintained and their difference gives us the time taken for search execution .