Search Engine Project

Team 19

Algorithms :

1- DFS:

After getting seeds , each threads starts crawling in DFS with one seed , by getting first allowable URL in the content of its seeds. This URL becomes as a seed where thread makes all validations on that URL . If it is valid , then its content will be parsed and so on.

2- How checking Robots.txt is done :

Each fetched URL is checked before being fetched if it is allowed or disallowed by Robots.txt. This is done by getting the Robots.txt of the root of it . Filtering the content of the file ,then searching for the URL in the filtered data to determine the next behavior.

3-Saving State :

In this algorithm we mainly use the data base to save the current state of the crawler we have made a table for the compacted strings to avoid duplicates of URLs or content and made a table for the seeds to put the current URLs have been parsed .The main reads this data from database every time it runs . filling the vector of compacted words (which represents saved documents content in a smaller word) with data from compacted table to check before downloading any documents that its content is not stored before.

Seeds table represents current seeds always ,not only as a start .As any time could be then end of running so that it will be start for the next.

4-HashMap<String, Vector<Help_data>> Detials = new HashMap<>():

we used hashmap of string & vector and this vector of class that contain [number of document , TF].

5 -PorterStemmer stemmer = new PorterStemmer():

Stemmer, implementing the Porter Stemming Algorithm The Stemmer class transforms a word into its root form.