Cairo University

Faculty Of Engineering

Computer Engineering Department

Search Engine Project

Submitted by:

Khaled Osama Ibrahim
Omar Said Labib
Omar Sayed Hassan

2017

Algorithms used:-

Web Crawler:

The main algorithm that builds the web crawler is breadth first search (BFS), it depends on the seeds we enter, in order to get their children "urls" as the first level seeds and then get their children and so on....
The connection to the url and the downloading of their html is done using "Jsoup" library.
The results that match specific conditions are synchronized among the threads using Database.

Indexer:

Libraries used: "google guava" for tokenization, "Portstemmer" for stemming words

The indexer stats indexing page right after the crawler is done, by accessing "indexerurl" table which contains the ids of html field that needs to be indexed fetchs the whole table and loops on all the ids.

For each id the page is parsed (the title and the body), remove stop words,tokienze and stemm.

For each word in the document its inserted in the "wordcnt" table the record contains the word itself, its position,boolean wether it was in the title or not .

.