Search Engine Flow Chart

Indexer

Start

Grab a URL from the crawled URLs list

Open Jsoup Connection and get the html document

Get all the text from the document

Remove the special characters and the stop words

Tokenizing the text

Calculating the TF of each word

Adding each word posting to its posting list

Are there any URLs left
— Yes
— No

Generate the database documents

Upload the documents to the database.

End

Crawler

Run

Crawl the seeds

Here we put them in the (URLsToCrawl) Queue and then we start Crawling

CrawlerNum less than 6000
— No
— Yes

Pick one URL from the queue

Open Jsoup Connection and Generate all URLs inside this URL

for each URL

Not visited URL & Crawled Num less than 6000 & Not Disallowed URL
— No
— Yes

Generate Disallowed URLs from the robots.txt

Update the VisitNum map & add this URL to the (URLsToCrawl) Queue& Increment CrawlerNum&store the state in the database

Ranker

Start

Grab a word from the query

Doc=Grab a relevant document

Get its TF & IDF

Get its PageRank score

Is the word in the URL
— Yes
— No

TF_IDF*=100

Is the word in a special position in the document

TF_IDF*=WeightOfPosition

Score =0.8TF*IDF+0.2*PageRank

Doc.FinalScore+=Score

Are there any documents left for that word
— Yes
— No

Are there words left?
— Yes
— No

Sort all the retrieved documents in descending order based on the final score

Return the documents back to the servlet

End