Search Engine

Crawler:

1. We start from seed set download each document and extract the links and extract the main tags of html
2. Add the links to non-crawled table in the data base and delete crawled URL from non-crawled and add it to crawled
3. We update page rank of each URL to use it in ranking by popularity
4. Avoid links that are forbidden on robot.txt

Indexer:

We get the links from the crawled URLs and get the main tags of html and extract the words then stem it and remove stop words then add to mongo dB database and store the values of TF and DF to use it in the rank by relevance

Ranker:

Rank by relevance :

1. We get the TF and DF from the database stored by the indexer
2. We calculate TF-IDF for web page by summing up the TF-IDF for query words by equation : $\log\left(\frac{indexedCount}{DF}\right) * \sqrt{TF}$
3. We check the importance of the query words by multiplying each sector as if mentioned in title or body or description or headers by certain weight such as : $\text{title} * 0.8 + \text{h1} * 0.7 + \cdots$
4. Then merge the 2 equations mentioned in(2) and in(3) by summing them

Rank by popularity :

1.we get the rank of each as mentioned above In crawler then merge the two factors (Popularity, TF-IDF) by the equation

$$rank = \frac{popularity * TF\_IDF}{popularity + TF\_IDF}$$

2. then sort the links by this rank

Query processor :

1. stem the words of the query and remove stop words
2. get the links of each word from the mongo dB

Phrase searching :

1. we extract the phrases either one phrase or multiple ones
2. we get the links by query processor and check over the links if the links contains the phrase exactly

Interface :

1. get query suggestions from database
2. receive the new query from the user and insert it In database
3. call the main engine that contain query processor and phrase searching and get the links
4. loop over the links and get some snippet from the body and then bold the query words in it and get the title
5. display the results