Search Engine Report

Team 11

# Algorithms Used

## Crawler

* + Reads seed or previous links from previous crawls and compact strings to complete a crawl
  + Parse valid Links for links inside them and storing the page compact string as not to parse the same page twice and normalizes links before storing them in a text file
  + If a link has a robots.txt that doesn’t allow parsing it won’t be parsed
  + It uses multiple threads to crawl and synchronized on shared variables

## Indexer

* + Reads Links to be indexed and reads previously indexed links
  + Stores tf if each word in a certain webpage and stores df, tags of each word removing stop words.

## Ranker

**Splitted into offline and online**

* + Offline: used pagerank algorithm to calculate popularity scores of pages. Also, the tf-idf is computed offline and stored in the database.
  + Online: Retrieves tf-idf scores and popularity scores. Then it maps the score of every relevant url and updating it by giving different weight for each tag and aggregating it with the score and then sorting the webpages according to the scores.It finally loops on the webpages’ bodies to find a relevant paragraph and sets the paragraph of each page

## Query Processor

* + Stem the query words by using a porter stemmer.
  + Retrieve the webpages that contains the stemmed words and map it to a hashmap to store both tags and urls.

## Servlet

* + The overridden doGet function takes partial query and retrieves queries from database that has the same prefix as the partial query. It stores queries in a database.
  + It writes the html and divide results to several pages and sets the bold words in each paragraph.