SERACH ENGINE PROJECT

Phase I - Report

Bassel Akmal Mohamed Kamal	1142027
Kerelos Diaa Shoukry	1142128
Fady Nasser Fawzy	1142158

APRIL 3, 2018

Libraries

Jsoup

This library is an open-source JAVA HTML parser that we used to parse and download HTML documents \ webpages.

MySQL Connector

This library is developed by Oracle which is necessary to establish a connection between MySQL server and JAVA.

BoneCP

This library is a free open-source JAVA library for Database Connection Pool which we used to manage and handle multiple simultaneous connections from different threads to the database.

Guava

This is a google core library in JAVA, which is needed by BoneCP to function correctly.

SLF4J

These are logging libraries required by BoneCP to function correctly.

Algorithms

Porter Stemmer

This algorithm is used for words stemming. During Indexing phase, before storing the tokens (words) found in a page, they are stemmed in order to ease the Query Processing and Searching phases.

The original stemming algorithm paper was written in 1979 in the Computer Laboratory, Cambridge (England), as part of a larger IR project, and appeared as Chapter 6 of the final project report.

The algorithm can be found in this link: https://tartarus.org/martin/PorterStemmer/

We modified the algorithm functions to match the requirements of our project.

Assumptions

- The crawler will crawl a fixed number of documents each iteration which is *iterationMax* (set to 2500), then based on a **selection criterion** it will re-crawl some important documents in the already-crawled (2500) documents.
- The **Selection Criterion** is setting a priority to some webpages according to the number of outlinks (links) found on that webpage, if this number exceeded *HighPriorityLinks* (set to 50), this webpage would have a high priority flag (*highPriority)* set to 1, and would be re-considered in the re-crawling process.
- The crawler will crawl a maximum number of pages *totalMax* (set to 5000).
- Pages that are fetched by the crawler are stored in a folder called “pages”, so as to be used in further phases such as: the indexer.