COMP 4047 Internet and World Wide Web

Group Project: Design and Implementation of a Search Engine

1. Project Specification

In this project, you will design and implement a search engine. It gathers information (keywords and URLs) from the Internet and then serves users' requests.

1.1 Gathering Information

Write a Java program to gather keywords in English from HTML documents and their corresponding URLs.

Data structures:

1. URL Pool and Processed URL Pool are used to store URLs, where URL Pool can store at most *X* URLs.
2. Design suitable tables to efficiently store the keywords and their corresponding URLs.

Algorithm:

1. Input and initialization: The user is prompted to provide: i) the URL of the web page which serves as the starting point of web search, and ii) the values of the parameters *X* and *Y* (where *Y* is used in Step 4). Assign this URL to URL Pool and set Processed URL Pool to empty.
2. Retrieve and remove a URL from URL Pool, add this URL to Processed URL Pool, and get the corresponding web page.
3. Process the web page obtained in Step 2 as follows:
   3.1 Extract all the keywords from this web page, where a keyword is a word that has at least three alphabets and does not appear in the following **stop list:**

   and, the, for, did, does, are, was, were, has, have, had, that, this, these, which, whose, who, whom, what, why, she, they, them

   For each keyword, store the following data items in some tables: i) the keyword, ii) URL, and iii) keyword position number (e.g., if this keyword is the 58th word on this web page, its position number is 58).

   3.2 Extract the URLs from this web page. For each of these URLs, add it to the URL Pool if it satisfies three conditions: i) it does not appear in URL Pool, ii) it does not appear in the Processed URL Pool, and iii) the number of URLs in the URL Pool is less than *X*.
4. If the number of URLs in the Processed URL Pool is less than *Y*, then go to Step 2; otherwise, stop.

1.2 Serving Requests

Set up a web server called *Abyss Web Server* such that it supports CGI and Java interpreter. You can download this web server from:

http://www.aprelium.com/

You may need ActivePerl and you can download it from:

http://www.activestate.com/activeperl

Write a Java program to serve users' requests and support the following:

1. **Keyword Matching:** The user's query contains one keyword. The program finds the URLs of the web pages which contain this keyword.
2. **Phrase Matching:** The user's query contains a phrase with two or more keywords. The program finds the URLs of the web pages which contain the given phrase.

The program composes a web page to list the above URLs. Then the web server sends this page to the user.

2. Submission and Demonstration

1. Forming Groups

Each group has three students. Form your own group and email the names and student IDs of your group members to Mr Yi Peipei at csppyi@comp.hkbu.edu.hk **on or before 21 September 2016**. If we do not receive your group information by this deadline, we will form a group for you. The finalized grouping will be posted on the course web page by **23 September 2016**.

2. Submission and Demonstration

2.1 Submit a ***hardcopy report*** that: i) describes the details of the design and implementation of your search engine, and ii) includes a signed ***Participation Form*** which is available on the course web page. Put this report into Prof. Y. W. Leung's mailbox **before 11:00pm on 1 November 2016 (Tuesday)**.

2.2 Prepare the following files:

- Program files (source files, executable files, CGI files, HTML interface, etc.).
- Data file which is obtained by executing your search engine with X=10, Y=100 and the following starting web page

  http://www.hkbu.edu.hk/eng/main/index.jsp

  The data file contains the gathered keywords and their corresponding URLs.

Each group (say, group x) packs the above files into one file called ***group_x.zip***, and submit this file to the HKBU Moodle **before 11:00pm on 1 November 2016**.

2.3 Demonstrate your search engine and explain its source code during **2 – 4 November 2016**. Mr Yi will arrange a suitable time slot and inform you. If you cannot explain your source code in the demonstration, it will be regarded as "suspected plagiarism" and it will be reported to the Department.

Assessment Criteria

1. Your must use Java to implement the specification given in Section 1.

2. *Design:* There are many alternatives to design a search engine. Your design will be assessed in several aspects:
   (i) Performance: Is it fast? Is it storage-efficient?
   (ii) Program structure: Is it easy to understand, maintain and modify your programs?

3. *Implementation:* Your implementation will be assessed in two aspects: i) Does it implement all the specified functions? ii) Are there bugs? iii) Is the implementation efficient?

4. *Documentation:* Your source programs should contain clear and useful comments. Your report should be clearly written and contain sufficient details.

5. *Marking scheme:* The marking scheme is specified in the participation forms. In addition, cooperation is very important and all group members must evenly share the work. This is one of the assessment criteria. If your group members could not share the work evenly for any reasons (e.g., a member is ill for a long period), please inform Prof. Y. W. Leung as soon as possible or there may be mark deduction.