Getting XML-based Information from a Remote Server through the Internet

Objective

Gain experience in getting information from a remote server through the Internet.

What to do?

As the Internet is filled with all kinds of information, we rely more on the Internet than ever. However, when information we required is quite large, how to get information in **an automatic and efficient** way become a question. As we know, information on the Internet is stored on millions of remote severs. When we request information, we have to send a request to these severs which will reply us with some information we need. So it is very useful for us to learn how to send requests to a remote server and get information we need from the server automatically and efficiently.

In this assignment, the students will develop a tool, which can send requests to a remote server automatically, and then extract useful information from the returned XML file by the remote server. The programming component of this assignment consists of four parts: (1) reading request information from a XML file; (2) sending requests to a remote server through the user API (Application Programming Interface) provided by the server; (3) getting the returned XML files from the remote server, and then extracting information; (4) writing the useful information received from the remote server to a local XML file; (5) displaying the received information on the browser with formatted style.

The PubMed (http://www.ncbi.nlm.nih.gov/pubmed) is a digital library in biomedical domain, which contains millions of biomedical literature. We can search information of biomedical literatures stored on the PubMed server using its API. In this assignment, the students firstly need to send some biomedical literatures' title to the PubMed server, and then obtain these biomedical literatures' document IDs (PMID) and other information by parsing the XML files returned by the server. Finally, the students need to write the received document information into a local XML-based file.

A XML-based file called "ITEC4020-A1-dataset.txt" which contains biomedical literatures' title will be provided for this assignment. The PubMed server user API (ESearch) is available. Please refer its documentations and examples at: http://www.ncbi.nlm.nih.gov/books/NBK25499/#chapter4.ESearch. Information returned by the PubMed server will be in XML format, which contains the document ID (PMID) and some other related information of the required biomedical literature. Please note that you need to extract the PMIDs first and then write these PMIDs into a XML file on your local machine. The PubMed may return to the unique PMID, multiple PMID and no PMID. Your results files should include all the documents found in PubMed, document not found in PubMed. The results file should be called "groupID_result" in XML format. The results file will be displayed in your browser with certain formatted style with DOCNO, literature title, authors, PMID(if found), publisher (optional) and published year (optional).

Useful links:

- Entrez Programming Utilities Help**:** http://www.ncbi.nlm.nih.gov/books/NBK25499/
- Sample applications: http://www.ncbi.nlm.nih.gov/books/NBK25498/?report=reader
- Introduction to Utilities (special characters) p3: http://www.ncbi.nlm.nih.gov/books/NBK25497/?report=reader
- E-utilitis p4: http://www.ncbi.nlm.nih.gov/books/NBK25499/?report=reader#!po=2.63158
- Search fields: http://www.ncbi.nlm.nih.gov/books/NBK3827/?report=reader#pubmedhelp.Search_Field_Descrip

What to submit?

You should submit the following items:

1. The assignment report that describes your methods and the tool that your group designs and implements.
2. The experimental result file, namely the "groupID_result" and style sheet file(if applied).
3. The source code.
4. A file called "readme.txt" where you give a tutorial on how to compile and run your programs.

How will you be graded?

The following will play a crucial role in your grade for this assignment.

1. Correctness of your program in obtaining PMID. An evaluation tool will be used to calculate the accuracy of your "groupID_result" file.
2. Your assignment report that should include introduction, description of your methods, description of your implementation in particular about how to solve and address the above problems, and analysis of the results.
3. Your group class presentation for the project.
4. Clarity of your programs (comments!).
5. Ease of using the README to test your programs.
6. Your collaborations with your team members.

The full mark for this assignment is 25. Your programs and assignment report account for 15 marks. Your team presentation accounts for 5 marks. Your participation from your team members accounts for 5 marks.