Retrieving and Analyzing XML-based Documents from a Remote PubMed Server

Objective: Gain experience in retrieving and analyzing 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. It is obvious that Internet can bring us convenience and colorful life. 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 be required to 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 an XML file; (2) sending requests to a remote server through the user API (Application Programming Interface) provided by the PubMed server; (3) getting the returned XML files from the remote server, and then analyzing these files and extracting information; (4) writing the useful information received from the remote server to a local XML file.

The PubMed (http://www.ncbi.nlm.nih.gov/sites/entrez/) 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) by parsing the XML files returned by the server. At last, the students need to write the document IDs into a local XML-based file.

A XML-based file called "4020a1-datasets.rar" which contains biomedical literatures' title will be provided for this assignment. The students first have to parse this XML-based file and then extract titles, and finally send these titles to the PubMed server. The PubMed server user API (ESearch) is available. Please refer its documentations and examples at: https://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 file should be called "groupID_result". The format of this file can be found in the file "group0_result".

What to submit?

You should submit the following items for your first assignment:

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", which ID stands for your group number (such as "6").
3. The source code and 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 document ID. An evaluation tool will be used to calculate the accuracy of your "groupID_result" file (such as "group6_result" for group 6) according to the golden standard from PubMed.
2. Your group 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 (please provide your comments!).
5. Time and space complexity of your programs.
6. Ease of using the README to test your programs.
7. 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. The group-peer marking from your team members accounts for 5 marks.