CS69012 Computing Laboratory II

Assignment 3

This assignment introduces web crawling, automatic extraction of data based on some grammar rules, loading of data into a database and querying information from the database.

The assignment is divided into the following 4 parts.

PART - 1 (Web Crawling):

Write a program to crawl the web pages of Department of Computer Science & Engineering, IIT Kharagpur faculty. You may access the CSE webpage by using the following link:

http://www.iitkgp.ac.in/department/CS

Crawl, download and save all the available faculty web pages. You can use python for this part, but please do not use a python application downloaded from the web. Write the crawl program on your own

PART - 2 (Creating a grammar):

Study the syntax of the HTML files downloaded in Part-1 and create a grammar that can be used to extract the following fields for each faculty member. Note that all other irrelevant fields must be ignored by the grammar.
Name of faculty:
Designation:
Responsibilities:
Phone no:
Email Id:
Research areas:
Website address:
Awards & Accolades with various fields such as award title, Year & etc:
Selected publications details with various fields such as title, authors, journal name, year & etc:

Current project details with various fields such as title, sponsoring agency & etc:
Group members details with various fields such as name, area of research & etc:

Submit the grammar in **Backus–Naur form** or **Backus normal form (**BNF).

PART - 3 (Parsing files to populate a database):

Develop a parser using Lex / Yacc (or similar tools such as ANTLR) to extract the above specified fields from the downloaded files. Design suitable database tables for storing the extracted data. Write SQL script to create required table structures and insert the required data into MYSQL database. Your database may contain a number of tables for effective and efficient management of data.

PART - 4 (Extracting Information from a database):

This part is based on client-server programming. On the client side, design a web interface to connect & query the database and present the result. On client side you can use client side web technologies such as HTML, JavaScript & AJAX. On server side, you can use JSP, PHP & Servlets for handling client request, database querying and processing of the result. At least the following queries should be supported by your application. For each query design the suitable interfaces on the client side. For example, for query (c), you must provide interfaces for entering the name of faculty and year of publication. The queries are as follows:

a) Given a responsibility i.e. chairman, GATE/JAM, find who is responsible for that, if any.
b) Given an award (e.g. young scientist), list the faculty members who have got that award, if any.
c) Given a name of faculty (e.g., Rajat) and a year (e.g., 2014), list all publications in that year from the unit.
d) Find which faculty who has the maximum number of phd students.
e) Find which faculty who has the maximum number of publications in a given year.
f) Find which faculty has the maximum number of projects.
g) Create an interface to enter any SQL query as String, execute it and present its result.

Deliverables:

1. A program for point 1 written in Python language.
2. A Text file containing the grammar in BNF.
3. The program/s based on Lex/Yacc(or similar tool) that use your grammar for extracting data and load the data into the suitable database.
4. The client-server programs developed in a suitable programming. It should provide interfaces for connecting to your database, querying the database and showing the result of query.

Marking Scheme:

Part – 1	-	10 Marks
Part - 2	-	5 Marks
Part – 3		
Lex/Yacc based programs	-	18 Marks
Table Design	-	2 Marks
SQL Script	-	5 Marks
Part - 4		
Overall GUI Design	-	5 Marks
Queries a to c	-	3 Marks each
Queries d to g	-	4 Marks each
Viva Voice	-	10 Marks
Coding Style	-	5 Marks
Design document (Block Diagram)	-	5 Marks
Test Case, Test Report	–	10 Marks

Deadlines:

February 8: Deliverable – 1, 2 & 3
February 13: Deliverable - 4

References:

1. https://en.wikipedia.org/wiki/Backus%E2%80%93Naur_form
2. **Lex & Yacc Tutorial – ePaperPress**
3. http://aquamentus.com/flex_bison.html