Design Laboratory (CS69202)

Spring Semester 2023

Assignment 3: Web Crawling, Extracting Information

Assignment date: January 10, 2024 and January 12, 2024

Important Instructions:

1. You need to design a menu-driven program to resolve user queries. We leave the design of the menu up to you, but keep in mind that your menu should be easy to use and address all queries. The user should also be able to go back to the previous menu.
2. Write Python code using PLY to extract the above fields. Your program should show all the possible query fields a user can ask for (from the above list items).
3. You must think correctly about what kind of errors can come in the process and try to handle them. Use the PLY package in python. PLY ref: https://www.dabeaz.com/ply/
4. You must NOT use any other parsing tools apart from PLY (ex: Beautiful Soup is a strict no or any other framework) . Should anyone not adhere to this instruction, they will be awarded ZERO marks.
5. Your code should address the objectives using PLY. Anyone found addressing the objective with no such use of PLY will be awarded ZERO marks.
6. Not adhering to these instructions can incur a penalty (worst case being 0 marks).
7. You can write a readme file to provide any particular instructions related to program execution steps, input format, or anything that you might think is useful for the evaluator while evaluating the assignment.
8. Create a log file to keep track of all input queries
9. Plagiarism in any form is not allowed. Students found copying/sharing code will be awarded 0 marks. You may discuss ideas, share your logic etc but you must not share/copy code at all costs.
10. All errors should be handled properly.
11. In case you make any design assumptions/choice, write a report along with the codes clearly stating the reason for your choice.
12. Submit the Assignments in <Roll>_CL2_A3.py and task as <Roll>_CL2_TS.py
13. Also submit a README file.
14. Save this in a folder named in the format: <Roll No.>_CL2_A3. Compress this folder to zip format, creating a compressed file <Roll No.>_CL2_A3.zip. Upload this compressed file to moodle. Example: If your roll no. is 22CS60R05, the folder should be 22CS60R05_CL2_A3, and the compressed file should be 22CS60R05_CL2_A3.zip.
15. Not adhering to these instructions can incur a penalty.

Day - 1 (10/01/24)

This assignment is on crawling web pages and extracting the required information by creating suitable grammar rules.
Assignment (Crawling ICC World Test Championship Wikipedia website→ https://en.wikipedia.org/wiki/ICC_World_Test_Championship)
This page contains all the essential information related to the wtc 2023-25 cycle, like teams and player details, venue, match results etc.

1. From the Wikipedia page for the 2023-2025 Tournament -
    - Given the names of two countries give the statistics of their match along with the stadium name and location where the match was held; If the match is yet to occur give only the stadium name and location and print “Yet to take place”
    - Given the dynamic nature of the page give a provision to periodically scrap the page and check for updates in case the match given in the query is yet to take place
2. Given a player name from the current squad.
    - Show his DoB
    - National Side
    - Role (Top-order Batsman, Middle-order Batsman, etc..)
    - Domestic clubs played for
    - Career stats (Matches, Runs scored, batting average, 100’s/50’s, balls bowled, wickets, bowling average, 5 wickets in an innings, 10 wickets in a match, best bowling, catches/stumping) in:-
        - T20
        - ODI (One Day Internationals)
        - Tests