Design Laboratory (CS69202)
Spring Semester 2025

Test: WebScrapping
Test date: January 20, 2025
Time: 2.15 PM - 4.30 PM
Total Marks: 100

Important Instructions:

1. You have to strictly use Python for this assignment.
2. You have to use BeautifulSoup and Selenium for this assignment. Use SQLite3 for the database purpose.
3. Not adhering to these instructions can incur a penalty (worst case being 0 marks).
4. 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 test.
5. Plagiarism in any form is not allowed. Students found copying/sharing code will be awarded 0 marks. You must not share/copy code at all costs.
6. All errors should be handled properly.
7. In case you make any design assumptions/choices, write a README along with the codes clearly stating the reason for your choice.
8. Submit the zip folder you received on your desktop. Put your roll number and system number in the proper place. ***<rollno>_<sysno>_WEBSCRAPPING_TEST***
9. Make sure there are 2 folders inside the parent folder. Folder Q1 should have *<rollno>_Q1.py* and a database file named *"UCL.db"*. Folder Q2 should have *<rollno>_Q2.py* and a database file named "*ws_sln.db"*
10. Write all the code in a single Python file for each question.
11. Not adhering to these instructions can incur a penalty.

Q1. Part - (a) [30]

The webpages of the UEFA Champions League are provided for the last 14 years (2010-11 onwards up to 2023-24) in .html format. On each page, there is a table of "Top goalscorers" given as follows:

Top goalscorers [edit]

Rank[35]	Player	Team	Goals	Minutes played
1	Karim Benzema	Real Madrid	15	1106
2	Robert Lewandowski	Bayern Munich	13	876
3	Sébastien Haller	Ajax	11	668
4	Mohamed Salah	Liverpool	8	1008
5	Christopher Nkunku	RB Leipzig	7	531
	Riyad Mahrez	Manchester City		986
7	Cristiano Ronaldo	Manchester United	6	611
	Darwin Núñez	Benfica		613
	Kylian Mbappé	Paris Saint-Germain		673
	Leroy Sané	Bayern Munich		798
	Arnaut Danjuma	Villarreal		906

You have to scrap this above table and store it in a database called "*UCL.db*". Create a table named "UCL_top_goalscorers" with the following columns:

a. Year ("2010-11", "2011-12" … so on)
b. Rank
c. Player
d. Team
e. Goals
f. Minutes played.
g. Additional counter as 1,2,.... [keep it as PK]

Make sure all the columns have valid entries for every row in the table.

2023–24 UEFA Champions League

Wembley Stadium in London hosted the final

Tournament details	
Dates	*Qualifying:* 27 June – 30 August 2023 *Competition proper:* 19 September 2023 – 1 June 2024
Teams	*Competition proper:* 32 *Total:* 78 (from 53 associations)
Final positions	
Champions	Real Madrid (15th title)
Runners-up	Borussia Dortmund
Tournament statistics	
Matches played	125

For each of the years ("2010-11", "2011-12" and so on), from the same webpage, capture the tournament details and store the following details in a separate table named "UCL_tournament_details":

a. Year ("2010-11", "2011-12" … so on)
b. Champions
c. Runners up
d. Matches played

Q1. Part - (b) [4+ 5 + 6 + 7 + (3+5)= 30]

Now answer the following questions based on the data stored in part (a) in "*UCL.db*"

QUERY 1) In which years, between the 2010-11 and 2023-24 seasons, did the number of top goal scorers exceed 13? (Print the years)

QUERY 2) Which club has finished as the runner-up the most number of times? [If there are more than one club, you may print any one of them]

QUERY 3) Which player(s) from the list of top goal scorers have played more than 1000 minutes but scored fewer than 6 goals in a particular season? [Print the name of all the players if there are multiple players.]

QUERY 4) How many distinct players from "Liverpool" have appeared in the list of top goal scorers across all seasons from 2010-11 to 2023-24?

QUERY 5) Which club won the most Champions League titles from the 2012-13 to 2017-18 seasons, (including both seasons)?How many times? Additionally, name the players who were the sole top goal scorer in each of these six seasons along with their no of goals scored.

Q2. [40]

You are provided with a webpage .html file

Your task is to perform the following steps using Selenium:

- Scrape the Table Data:
    - Navigate through all the pages of the website until the end.
    - At the last page, there is a table containing some form of data. Extract the data from the table displayed on the webpage.
    - Store the scraped data into an SQLite database named ***ws_sln.db***.
    - The database should contain a table named *words_10* with the following columns:
        - ID (unique identifier for each entry)
        - Word (the word from the table)
        - Length (length of the word)

- Repeat the above steps 10 times, storing the scraped data into the same database table each time. Print the Total Number of words stored in the database.