Assignment 2

Web Crawling and Extracting Information

Task 1 :
(Crawling 2020 Summer Olympics Wikipedia website :
https://en.wikipedia.org/wiki/2020_Summer_Olympics)

This page contains all the essential information related to the 2020 Summer Olympics.

- Write a python code that reads a URL & saves the page in a file in HTML format.
- Extract all the teams and cities that participate in the 2020 **Summer Olympics host city election.**

  ==

Task 2: (Creating grammar and parsing the files)

Create grammar that can be used to extract the following fields:

a. For 2020 Summer Olympics extract:
  i)Host city
  ii)Motto
  iii)Nations
  iv)Athletes
  v)Events
  vi)Opening
  vii)Closing

b. 2020 Summer Olympic Sports program:-
  i.For any five sports extract:
    Venue
    Dates
    Competitors
    Number of events

c.Medal table(for Top 10 countries):-
  i) Rank
  ii) Number of Gold , Silver and Bronze
  iii)For any given country name out of top 10 countries in medal tally print:
    i) Flag Bearer(opening)
    ii)Flag Bearer (closing)

iii) Multiple Medallist Names and their sport(if any)

iv) Given two country names(only for top 10 countries in medal tally):

a. Compare their medal tally

INSTRUCTIONS

1. You can ignore other fields except the above.
2. 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.
3. 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).
4. 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/
5. 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.
6. Your code should address the objectives using **PLY.** Anyone found addressing the objective with no such use of **PLY** will be awarded **ZERO** marks.
7. Not adhering to these instructions can **incur** a penalty (worst case being **0 marks**).
8. 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.
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.

DELIVERABLES

Submit all the python files in a folder named in the format: <Roll No.>_DesLab_A2. Compress this folder to zip format, creating a compressed file <Roll No.>_DesLab_A2.zip .Upload this compressed file to moodle.