Assignment 8
Web Crawling and Extracting Information-Part1

Computing Lab-II
12th Mar 2021

This assignment is on crawling web pages and extracting the required information from them by creating suitable grammar rules.

Task 1 (Crawling RottenTomatoes website)

1. RottenTomatoes is an IMDb like website, where you can find an online database of information related to films, television programs, including cast, production crew, personal biographies, plot summaries, trivia, ratings, critic and fan reviews.
2. You are provided with a file named "rotten tomatoes movie genre link.txt," which contains URL links for ten different genre-wise top 100 movie lists.
3. Write a python code that reads each of the URLs, saves the pages in HTML format.
4. Now given a user input of any of the ten genres, you should list all the movies in that list and wait for user input of a particular movie name from the list.
5. Given a movie name as the input, you should download and save the corresponding movie page's HTML file.

Refer: *https://programminghistorian.org/en/lessons/working-with-web-pages*

Task 2 (Creating grammar and parsing the files)

1. After saving movie pages in HTML format, try to study the syntax of HTML files.
2. Create grammar that can be used to extract the following fields for the movies.
   - Movie Name
   - Director
   - Writers
   - Producer
   - Original Language
   - Cast with the character name
   - Storyline
   - Box Office Collection
   - Runtime
3. You can ignore other fields except the above.
4. Write (python code using PLY) or (C code using Lex,Yacc) to extract the above fields. Your program should show all the possible query fields a user can ask for(from the above list items). And according to the user selection, it should show the corresponding field for the particular movie.
5. Your program should also save the result in a log file as per the following format.
   <Genre> <Movie_name> <Field_requested> <Field_value>
   For a field with multiple values, it should make an entry for each value.
6. You have to think correctly about what kind of errors can come in the process and try to handle them. Note that you can not use the "Beautiful Soup" python package for this assignment. **Use the ply package in python / Or you can code in C using lex and yacc.**
   Refer: http://www.dabeaz.com/ply/

Deliverables: codes for task1 and task2

Evaluation Scheme

Task1: 15 marks
Task2: 65 marks (all the fields grammar + correct output)
Error handling: 10 marks
Coding Style: 10 marks

Important Instructions

1. Plagiarism Rule: If your code matches (more than 50%) with another student's code, all those students whose codes match will be awarded zero marks without any evaluation. Therefore, it is your responsibility to ensure that neither you copy anyone's code nor anyone can copy yours.
2. Code error: If your code doesn't run or gives an error while running, you will be awarded zero marks.