Assignment 5

Web Crawling and Extracting Information

Computing Lab (II)
28th Jan 2020

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

Task 1 (Crawling DBLP profiles of Indian CS professors)

1. DBLP is a computer science bibliography website which contains reference for bibliographic information on major computer science publications.
2. You are provided with a file named "**professors.txt**" which contains the dblp profile urls of some of the renowned CS professors at premier institutes in India. Each line in the file contains the url of one professor
3. Write a python code which reads each of the urls, saves the profile pages in html format.

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

Task 2 (Creating grammar and parsing the files)

1. After saving all the profiles in html format, try to study the syntax of html files.
2. Create grammar that can be used to extract the following fields of all the publications.
    a. Profile name (professor's name i.e., the person whose profile you crawled)
    b. Title of the paper
    c. Type of publication (book/journal/conference/part_in_book/editorship/reference_work/informal)
    d. Venue of the paper
    e. Year of publication
3. You can ignore other fields except above.
4. Write python code using PLY to extract above fields. Your program should have the filtering functionalities by taking filter criteria for each field as input (described below).
    a. <text> in name (<text> as input)
    b. <text> in title (<text> as input)
    c. b/j/c/p/e/r/i (publication type as input)
    d. <text> in venue (<text> as input)
    e. <year> exact match (<year> as input)
5. All filtering must be done in case-insensitive way (consider uppercase and lowercase characters in the fields as same). Ignore all symbols in the fields (consider only letters, numbers, and spaces). Note that if filter criterion for any field is not entered or kept empty, then you program should not do any filtering on that field.
6. You have to think properly on what kind of errors can come in the process and try to handle them. Note that you can not use "Beautiful Soup" python package for this assignment. Use ply package in python.

7. After filtering (based on the entered filter criteria) from all the crawled html files, your programme should print the filtered publication entries in the following format.
   <name> <title> <pub_type> <venue> <year>

Refer: http://www.dabeaz.com/ply/

Deliverables: codes for task1 and task2

Evaluation Scheme
Task1: 10 marks
Task2: 50 marks (all the fields + proper filtering + correct format)
Error handling: 10 marks
Coding Style: 10 marks
Viva voce: 20 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 with zero marks (may be with -ve marks too depending on the situation) without any evaluation. Therefore, it is your responsibility to ensure that neither you copy anyone's code nor anyone is able to copy yours.
2. Code error: If your code doesn't run or gives error while running, you will be awarded with zero mark.