CECS 571 Fundamentals of Semantic Web Technologies

Spring 2021

Project 2: Semantic Data Generation

Overview

In teams of 4, visit data.gov to select 2 datasets of interest and write a program to convert these datasets to semantic standards in .rdf or .owl format. For example, your program may accept input in XML or CSV (or any other format) and output the converted information in RDF or OWL format.

It is advisable to choose datasets that can be related to one another in some way, e.g. one dataset may be traffic records of the LA area, and the other dataset may be event records of the city Long Beach. Once combined, we can ask interesting questions over these datasets such as “which days/time of the day are best to visit the CSULB campus with the least amount of traffic?” or “Did the 49ers game on Sunday generate more traffic than the usual?” etc., which will be the focus of Project 3. Therefore, the data you are generating in Project 2 is in view of being queried over in the subsequent Project 3.

You can choose any programming language (e.g. Java, Python) provided that your team is also comfortable with. You are free to select datasets in any existing format on data.gov (e.g. HTML, XML, CSV).

You are free to choose any domain that interests you, e.g. music, film, food, weather etc. When selecting domains, keep in mind that these chosen datasets should ideally help you answer interesting questions if combined (in the subsequent project 3). For example, you may have chosen a dataset about concerts held in the LA area last year, and another seemingly unrelated dataset about weather records for LA last year, but once both datasets are converted into semantic data, we could query these datasets with complex questions such as “how was the weather for the concert held in Hollywood Bowl on May 17, 2019?” in our efforts to demonstrate the power of Semantic Web technologies that is beyond keyword search.

Requirements

Your code must be hosted on GitHub and is self-contained, i.e. once downloaded, anyone can open and run your program without having to configure any other dependencies.

Deadline

Monday, Mar. 22, 2021, 5pm local time

Submission

Submit three items electronically to the designated DropBox folder on BeachBoard:

(1) Presentation slides in .pdf or .ppt summarizing how you approached this project. Presentations should be prepared for a 10-minute talk and 2-minute Q&A. In particular, your talk should include the following:
    a. Why did you choose these datasets?
    b. What kind of questions are you hoping to answer in the future?
    c. Code snippets of your program or any technical highlights you would like to share
    d. Technical challenges encountered and how you overcome these issues?
(2) The .rdf or .owl output generated from your program; and
(3) A link to your project's GitHub page in the comment field as you submit your presentation slides to BeachBoard.

Grading Guidelines

This assignment is worth 20% of the final grade. All members of the same team will receive the same points, which are subject to peer review (see below). The code/semantic output and the presentation are marked out of 10% each. You will be graded on the extent to which the required deliverables discussed above have been successfully met.

In particular, the semantic data generated from your program will be judged on its logic, consistency, class & relationship design, and the appropriateness of the relations & structures within. Furthermore, your code will be judged against conventions, complexity, reuse and extensibility. Finally, the presentation will be judged on the quality of the discussion and the overall clarity.

Peer review: contributions in a team-based assignment should be understood as the individual input that is valued by your peers and is advancing the collective team outcome positively. Thus, your final grades may be adjusted to reflect the evaluations rated by your peers. Please be reminded that the instructor cannot give credit to any individual who makes little or no contribution to group-based assignments. Further peer review submission instructions will be provided to the class after the due-date of this assignment.

Tips on Earning Points

Consider the following questions when completing this project:

- Have you chosen an appropriate domain that offers a sufficient platform allowing you to answer complex questions in the subsequent project 3?
- If the semantic datasets you have generated are ran in Protégé, can the reasoner pick up any errors or inconsistencies?
- How many different kinds of relationships are modeled in your semantic data?
- How complex is the semantic output overall?
- Can someone not involved in the design or development of your project gain a basic understanding of the various aspects of the semantic output after your presentation?