INF 558/CSCI 563: Building Knowledge Graphs

Homework 4: SPARQL

Released: Feb 7th, 2020
Due: Feb 16th, 2020 @ 23:59

Ground Rules

This homework must be done individually. You can ask others for help with the tools, however, the submitted homework has to be your own work.

Summary

In this homework, you will load RDF triples to a Triple-Store framework and then write SPARQL queries to retrieve data and perform simple analysis.
As for the Triple-Store, you will use Apache Jena, an open source lightweight Semantic Web framework that is easy to use and provides a programmatic environment for RDF, RDFS, OWL, and SPARQL. Additionally, you will run additional queries over Wikidata's endpoint.

Task 1: Apache Jena (6 points)

In this task, you will work with RDF triples over Apache Jena. First, download and install Apache Jena Fuseki (SPARQL server) from: https://jena.apache.org/download/index.cgi

Next, upload the attached graph data (triples) `lmdb.nt` to Apache Jena and write SPARQL queries to accomplish the following sub-tasks (1.1-1.4 are worth 1 point each, 1.5 is worth 2 points). Execute your queries on Apache Jena and include the results (screenshots) in your report. All queries should be limited to 10 results.

1.1. Get names of actors whose name starts with the letter 'B', and the names of films they starred in (total of 2 columns).

1.2. Get URIs and names of actors who are also producers but are not directors (total of 2 columns).

1.3. Get names of actors who starred in more than 20 films with the same director (who is not the actor). Show the name of the director and the number of films as well (total of 3 columns).

1.4. Get the names of countries, the number of drama films (films with genre "Drama") and the population (in millions) for each country. The results should be sorted by the number of drama films in descending order (total of 3 columns).
- Note that the population data in the graph is not in millions.

1.5. Get the names of female actors who were born after 1970 and their date of birth (total of 2 columns).
- Note that you do not have the data about gender and date-of-birth for actors in the provided graph. You will need to retrieve these values from an additional SPARQL endpoint using a federated query extension (the `SERVICE` keyword, read more about it here: https://www.w3.org/TR/sparql11-federated-query/).
- Hint: use DBPedia

* The attached file `model_example_instances.ttl` includes example instances ("model snapshot") from the graph to help you understand the model and write your queries.

Task 2: Wikidata (4 points)

In this task, you will use Wikidata's query service (SPARQL endpoint) to execute queries. The service is available here: https://query.wikidata.org/

Write SPARQL queries to accomplish the following sub-tasks (each is worth 2 points). Execute your queries in the GUI and include the results (screenshots) in your report. All queries should be limited to 10 results.

2.1. Get URIs and names of actors who are also musicians and were part of (P361) a band sometime between 1960 to 2010. Also, show their date of birth. If an actor doesn't have date of birth, the value can be null, empty string, or missing (total of 3 columns).

2.2. Get URIs and names of films that include at least 4 cast members (P161) who have been nominated after the year 2000 for the Academy Award for Best Actor (Q103916). Also, show the number of such actors for each film (total of 3 columns).

Submission Instructions

You must submit (via Blackboard) the following files/folders in a single `.zip` archive named `Firstname_Lastname_hw04.zip`:

- `Firstname_Lastname_hw04_report.pdf`: pdf file with your screenshots for Tasks 1 and 2
- Query files for Task 1:
  - `Firstname_Lastname_hw04_task_1_1.sparql`
  - `Firstname_Lastname_hw04_task_1_2.sparql`
  - `Firstname_Lastname_hw04_task_1_3.sparql`
  - `Firstname_Lastname_hw04_task_1_4.sparql`
  - `Firstname_Lastname_hw04_task_1_5.sparql`
- Query files for Task 2:
  - `Firstname_Lastname_hw04_task_2_1.sparql`
  - `Firstname_Lastname_hw04_task_2_2.sparql`