SOEN 363: Data Systems for Software Engineers

Assignment 3, Fall 2024

November 6, 2024

Date posted: Thursday, November 7th, 2024.

Date due: Sunday, December 1st, 2024, by 23:59.

Weight: 5% of the overall grade.

Individual assignment. You must work strictly on your own.

Overview

In this assignment, you create a NoSQL database of movies and their information. The movies data are directly extracted from assignment 2 and transferred into the NoSQL database.

Implementation Platform

We use Neo4J [1] in this assignment. While you may find many tutorials online, attending the tutorials sessions are strongly recommended. For any help re: programming, or questions on the platform, please see PODs.

Data Transfer

The data transfer is done by converting the data from each relation from the RDBMS into a csv (or tsv) or json data. You will then use the data and directly import it into Neo4j.

- `https://neo4j.com/developer/guide-import-csv/`
- `https://neo4j.com/labs/apoc/4.1/import/load-json/`

Entities / Nodes

In this assignment you creating the following entities (nodes) and populate the data.

- Movies (attributes: title, plot, content rating, viewer rating, release year, genres, languages, AKAs, and optional watchmode id)
- Actors (first name, last name)
- Countries
- Keywords

Data Files and Scripts

[10 pts] Extract the data from your database into the data file (csv, tsv, or json[1]).

[40 pts] Write scripts to create the database and populate the data in Neo4J.

Note that Neo4J supports array attributes, which are normally represented using weak entities in relational model:

`https://neo4j.com/docs/cypher-manual/current/functions/list/`.

To populate the such data (i.e. genres, languages), you may use a separate csv file.

[1]Using JSON is not recommended, but is permitted. Naturally, a relational data may be directly represented using a tabular data format such as CSV, TSV, etc.

Queries

Provide the answers to the following:

A) **[5 pts]** Find all movies that are played by a sample actor.

B) **[5 pts]** Find the number of movies with and without a watch-mode info.

C) **[5 pts]** Find all movies that are released after the year 2023 and has a viewer rating of at least 5.

D) **[5 pts]** Find all movies with two countries of your choice. Make sure your query returns more than one movie. List movies that may be associated with either of the countries (not necessarily both).

E) **[10 pts]** Find top 2 movies with largest number of keywords.

F) **[10 pts]** Find top 5 movies (ordered by rating) in a language of your choice.

G) **[5 pts]** Build full text search index to query movie plots.

H) **[5 pts]** Write a full text search query and search for some sample text of your choice.

Make sure all above queries return data. Modify the data in your database, if necessary.

Submit your assignment electronically on Moodle: `https://moodle.concordia.ca`
Include your name and student ID in the submission. Make sure that you upload the assignment to the correct assignment box on Moodle. No email submissions are accepted. Assignments uploaded to the wrong system, wrong folder, or submitted via email will be discarded and no resubmission will be allowed. Make sure you can access Moodle prior to the submission deadline. The deadline will not be extended.

References

1. `https://neo4j.com/`