Assignment 1: The Relational Data Model - Returned

Title	Assignment 1: The Relational Data Model
Student	Sannath Vemula (sannathreddyv)
Submitted Date	Sep 26, 2019 11:12 PM
Grade	**15.20 (max 16.00)**

Instructions

Objective
Let's play Data Science! You are provided some relational data and asked to illustrate some exciting insights hidden within it. You have two weeks, but you should demo a prototype visualisation to the whole class during the lecture prior to the due date. There is no pre-defined "correct solution", but you need to demonstrate mastery of the relational data model in the context of doing Data Science.

Data
The MovieLens dataset (https://grouplens.org/datasets/movielens/) is a publicly-released dataset of user movie ratings, built with a relational schema. There are four different datasets, and you are welcome to use any of them:
  (a) 20M (20 million ratings). This is the standard dataset, but you might find it quite large when just starting out with your assignment.
  (b) Small (100,000 ratings). This is a downsampling that might be useful for initial prototypes
  (c) Full (27 million ratings). This isn't really much larger than the 20M dataset; so, I don't recommend it.
  (d) Synthetic (1 billion ratings). If your primary interest is in Big Data, this might be fun to play with!

Feel free to *integrate* the MovieLens data with other sources (e.g., IMDB). This can definitely demonstrate an "extension of course expectations"; however, you are required in this assignment to use one of the MovieLens datasets as a principle element of your visualisation.

Implementation
It is expected that you will use Python, as it is the language of choice in Data Science. Moreover, it provides many visualisation libraries that will help you achieve more than you realise you can!

Pre-submission Demo
In class on 23 Sept, each group will have 3-5 minutes (depending on the number of groups) to demo their prototype to the entire class. You should have a functional visualisation by this stage, and focus the demo on something that you think will generate rich discussion. The discussion will be taken collectively *after the last group presents*; so, a more exciting demo will likely garner more peer feedback.

Submission
You should submit your implementation in raw python and a short technical report (3-5 pages) in pdf format. The markers will validate that the implementation works as described in the report. The report will describe:

- the insights revealed by the visualisation/application
- the design choices in the implementation/visualisation
- challenges encountered (if relevant)
- how this submission meets the requirements set out in the rubric (i.e., a justified self-evaluation)
- other details that you consider relevant

Grading

The entire group will receive a grade based on how well the submission adheres to the rubric below. Note that your report provides the opportunity to persuade the markers, but that they will ultimately grade according to the rubric.

The demo on 23 Sept will *not* form part of the grade; however, it is an opportunity to solicit peer and instructor feedback a few days before the deadline that may ultimately improve the quality of your submission (i.e., your grade). Also, having a prototype visualisation to present will contribute to your participation grade in the course.

Assignment 1 Rubric

Component	Weight	4	3	2
Transforming Raw Data to Insights	40 %	Visualisation is extremely informative. Unintuitive information is readily evident and requires no further explanation	Visualisation is informative, albeit possibly with some support from text	Visualisation is co Information is pres may lack interestir
Data Modelling	40 %	Demonstrates a clear mastery of the relational data model. Modelling of the data is flawless and/or demonstrates knowledge of both the strengths and limitations of the data model.	Schema is appropriate for the underlying dataset(s) and effectively addresses challenges with raw data	Relational schema fails to address lim the raw data.
Algorithmic Design	10 %	The project is clearly and demonstrably scalable. The visualisation may include a component that demonstrates scalability or the report may include a cost model	Algorithmic considerations are well reasoned; the visualisation has only minor performance problems	Some attempt is m address questions efficiency, but the visualisation is res small datasets
Relationship to Modern Practice	10 %	State-of-the-art considerations for the relational data model are clearly evident and add substantial value to the project.	State-of-the-art considerations for the relational data model are evident and add value to the project.	Some attempt is m engage with resear presented in class, limited value to the

Tips

 1) Try to load the data into a usable structure early (first few days). You may find it more difficult than you expect to model the raw datasets as workable data structures.

 2) Ascend quickly to a working visualisation and add complexity later; i.e., first build a minimal viable product (MVP). You will be exposed to new research and new ideas in class as you work on the assignment, so you want to be *agile* with your development patterns. At the same time, even in groups, it may take longer than you expect to go from raw data to a working visualisation; so, you don't want to leave this until the last few days before the demo or you could end up with nothing to show at all!

 3) Use libraries prolifically.

 4) Consider addressing each rubric component individually in your tech report.

Extra Resources

Over the course of the next two weeks, supplemental resources (e.g., instructions for Python libraries) may be added to the Resources/ panel of the course connex page.