2030ICT/7030ICT

Introduction to Big Data Analytics

Practice-based Assignment

Assignment 1

Trimester 2, 2021

Assignment 1

Description

Use the following two data files:

titles.csv - Contains the following information for titles:

- tconst (string, required) - alphanumeric unique identifier of the title
- titleType (string) – the type/format of the title (e.g., movie, short, tvseries, tvepisode, video, etc)
- primaryTitle (string, required) – the more popular title / the title used by the filmmakers on promotional materials at the point of release
- originalTitle (string) - original title, in the original language
- isAdult (boolean) - 0: non-adult title; 1: adult title
- startYear (number) – represents the release year of a title. In the case of TV Series, it is the series start year
- endYear (number) – TV Series end year. '\N' for all other title types
- runtimeMinutes (number) – primary runtime of the title, in minutes
- genres (string) – includes up to three genres associated with the title

ratings.csv – Contains the IMDb rating and votes information for titles

- tconst (string, required) - alphanumeric unique identifier of the title
- averageRating (number, required) – weighted average of all the individual user ratings
- numVotes (number) - number of votes the title has received

Tasks

1. Create a database named "imdb" on the Compass tool.
2. Create two collections named "titles" (refers to title.basics.tsv.gz) and "ratings" (refers to title.ratings.tsv.gz).
3. Schema validation: Write JSON Schemas for each of the collections based on the descriptions.
4. Import data to the collections using Compass. Insert the data in titles.csv into "titles" and ratings.csv into "ratings".
5. Perform schema analysis and describe the data characteristics. Go to the "Schema" tab and click on the "Schema analysis" button. Compass will generate an analysis for each column. You should describe the interesting characteristics of the data in the columns. You can learn more about schema analysis at: https://docs.mongodb.com/compass/master/schema/
6. Perform some advanced analysis of the data using Aggregation. You must extract the following information and include the output.
    a. Find the total number of movies released each year.
    b. Find the top five Fantasy-Adventure movie titles (primaryTitle**) released in 2021** according to the rating. [Hint: Genre must include both Fantasy and Adventure.]
    c. Find the top five Fantasy-Adventure movie titles (primaryTitle) **released in 2021** according to the number of votes. [Hint: Genre must include both Fantasy and Adventure.]

Marking Criteria

Database Creation and Import Data (Tasks 1-4) **10 Marks**	Database and collection creation	2 Marks
	Schema validation	6 Marks (3 Marks per schema)
	Import data	2 Marks (1 Mark per collection)
Schema Analysis (Task 5) Comment on the statistics of the column data, e.g., In which year we can find the greatest number of released movies? Which genre covers the highest number of movies? etc. **5 Marks**	Complete the assigned tasks using Compass	2 Marks
	Statistical analysis	3 Marks
NoSQL Queries (Task 6) **10 Marks**	Query 6.a Hint: $sortByCount	2 Marks
	Query 6.b Hint: $match, $lookup, $unwind, $project, $sort and $limit	4 Marks
	Query 6.c Hint: $match, $lookup, $unwind, $project, $sort and $limit	4 Marks
	Total	25 Marks

Good luck ☺