Test 2

DSA508: Big Data Platforms & Analytics

Problem 1

The sample_mflix database in MongoDB contains information about movies, users, and comments. It is ideal for exploring relationships between genres, ratings, and textual content. This problem emphasizes analytical storytelling and interpretation of insights derived from data.

a. Perform a complete exploratory data analysis using MongoDB Query Language (MQL) pipelines on the sample_mflix database. Summarize key descriptive statistics about movies, ratings, user engagement, and other relevant dimensions from all the collections available. Provide a clear storytelling of the EDA for this document database; you should go beyond just coding, a clear understanding of the context and the data is required.
b. Use aggregation pipelines to explore temporal or categorical patterns such as rating trends over years or by genre. Present clear visual or tabular summaries of your findings. Conduct a genre-based text analysis by extracting and analyzing phrases from movie descriptions, titles, and/or user comments to identify common linguistic or thematic patterns. Provide insights.
c. Apply topic modeling using a method such as latent Dirichlet allocation or non-negative matrix factorization to detect latent themes across movie descriptions or comment texts. Write a narrative interpretation of your findings, connecting the discovered topics and phrases to movie genres and audience trends.

The evaluation of problem one will emphasize the clarity and depth of your interpretation more than the coding complexity.

Note: the sample_mflix is also part of the sample databases loaded with sample_airbnb and week 9 assignment data.

Problem 2

In this exercise, you will combine your understanding of document databases in a cloud environment. You will upload and manage a document database on Azure using MongoDB or Cosmos DB. The data must be available online on Azure, while the dashboard can be executed locally using the appropriate connection information.

a. Prepare a document-based dataset and upload it to your Azure MongoDB or Cosmos DB environment. Ensure that the database and collections are properly structured and accessible online. You can use an existing document database as long as the dashboard and the analysis make sense.

b. Design and build a dashboard using Streamlit, or a similar tool connected to your online database. The dashboard should visualize important metrics and trends extracted from your document database.

c. Include a written narrative explaining the purpose of the dashboard, its analytical value, and how it can support decision-making in a realistic business or social context such as retail, education, or IoT.

Ensure that your submission includes both the database connection and the visual dashboard file, along with your explanatory narrative.