LSC Final Project

Group Name: user_lsc_25

Members: Umair X

Title: Exploratory Movies Annotator

Category (A/B/C):

A) APIs of the Spark ecosystems: advanced aspects of RDD/Dataframe/ SparkSQL/streaming Spark, libraries such as GraphX and Mllib.

Description:

- **Motivations and Goals**:

The Genre and rating play an important role in the interpretation, selection, and discovery of the movies. This allows for an analysis of how generic, rating, labels and tags are used, how genres and rating are formed and how these ingredients change, and how genres are compared according to rating, and how these data impact to people. In this modern era of technology, we can find easily our desired data through it. We use pySpark data frame techniques in a large collection of movies datasets to investigate important questions of genres, movies and rating.

Using some RDD Data frame techniques, we can provide a refine data to user. It will help user to experience his search easily by selecting specific genres and ratings so that user can easily entertain himself and find movies according to their specific category.

- **Technical Description:**

The proposed exercise base on the dataset of movies. There are three datasets which are interlinked with each other, which are movie, rating and tags. By using pySpark data frame we processed many queries to extract meaningful categorical output.

Programming Exercises (at least 5 Questions) with Solutions:

1. Extract the most rated top 5 movies?
2. Specify the number of movies according to the tags?
3. Categorize all funny and actions movies?
4. Find the top-rated movie according to genres?
5. The most romantic movie according to ratings?

Links:

/home/user_lsc_25/final_project/