CS322:Big Data

Final Class Project Report

Project (FPL Analytics / YACS coding): _____FPL Analytics____ Date: _01/12/20__

SNo	Name	SRN	Class/Section
1	Raghav Goyal	PES1201800111	D
2	Ishan Agarwal	PES1201800291	D
3	Akash Kumar Rao	PES1201800089	D
4	Mohit Gaggar	PES1201800112	J

Introduction

The project is based on real-life solutions to real time data analytics involved during sports live streaming. The project involves data analytics on streaming data and using that to provide suitable information according to the user needs. This kind of analytics happens at the backend of the sports app/websites. Since the data streamed is of huge amount, it requires us to use the big data concepts for fast and reliable use of the data.

Related work

- Hadoop The Definitive Guide by Tom White
- Pyspark official documentation : https://spark.apache.org/docs/latest/api/python/index.html

Design

Dataframe scheme for player details: stores all the information required for calculation of metrics and player profile

root

|-- name: string (nullable = true)

|-- birthArea: string (nullable = true)

|-- birthDate: string (nullable = true)

|-- foot: string (nullable = true)

|-- role: string (nullable = true)

|-- height: string (nullable = true)

|-- passportArea: string (nullable = true)

|-- weight: string (nullable = true)

|-- Id: string (nullable = true)

|-- no_of_fouls: integer (nullable = false)

|-- no_of_goals: integer (nullable = false)

|-- no_of_own_goals: integer (nullable = false)

|-- pass_accuracy: integer (nullable = false)

|-- shots_on_target: integer (nullable = false)

|-- no_of_matches_played: integer (nullable = false)

|-- rating: integer (nullable = false)

Dataframe schema for chemistries dataframe: stores the chemistries for each pair of players and is updated after every match

root

|-- _c0: string (nullable = true)

|-- player_1: string (nullable = true)

|-- player_2: string (nullable = true)

|-- chemistry: string (nullable = true)

The flow of program is:

- Streamed data is inputted from the socket at port 6100.
- Initializing the dataframes.
- Various metrics are calculated and stored in the dataframe for each event.
- At the end of the match chemistries are calculated between each player that played in the match.
-

Models Implemented:

- **K-Means clustering** is implemented using MLlib library which clusters the data in 5 clusters which is then used to get the mean of clusters and approximate the rating and chemistry.
- **Linear Regression for degree 2:** The regression algorithm is implemented using MLlib which takes age of the players and corresponding rating as training data and fit the model. The fitted model than predicts the rating by taking in the age of the player.

Results

The metrics are calculated using the data and the player profies are maintained in the dataframes.

Problems

- Setting up the spark required us to understand the spark properly and find the right tutorial.
- Streaming the data took us a while to understand how the flow will occur.

Conclusion

The working behind the live sports event is huge and involves large sums of data which requires careful handling and manipulation to get the desired and useful information to enrich the sports viewing experience.

EVALUATIONS:

SNo	Name	SRN	Contribution (Individual)
1	Raghav Goyal	PES1201800111	Chemistries calculation and handling
2	Ishan Agarwal	PES1201800291	Maintaining the player profile
3	Akash Kumar Rao	PES1201800089	Building the user queries accepting code
4	Mohit Gaggar	PES1201800112	Machine learning model building

(Leave this for the faculty)

Date	Evaluator	Comments	Score

CHECKLIST:

SNo	Item	Status
1.	Source code documented	
2.	Source code uploaded to GitHub – (access link for the same, to be added in status)	
3.	Instructions for building and running the code. Your code must be usable out of the box.	