

CS322:Big Data

Introduction

This is a streaming spark project which requires us to stream data from the given encoded dataset of matches and events. We then process it to find winning chances of a team. Winning chances are calculated by maintaining player profile which in turn is used to calculate player ratings and chemistry between players. Player profile and chemistry are maintained and updated for every batch of incoming data.

Related work

We have used StackOverflow extensively and official spark website.

Design

We start by creating a spark context which is used for creating streaming spark context. It is used for accepting the data from localhost using port 6100 and TCP connection. The data is streamed and processed every 5 seconds. We used transformation operations on the present DStream to get the required features. To get the features, we have used key value pairs as done in word count and used map reduce to calculate the desired player profile. The player profile has key as the player ID, and values as the dictionary of various required attributes. Next, we calculated the change in rating depending on the events which have key as player ID and value as change in rating. In similar fashion we calculated the new player ratings by updating the player profile with the previous values.

Chemistry between the players are calculated using the changed ratings which are used to create pairs of IDs and the value is found using the given formula. By using map reduce, we calculate the average of chemistries between the lineup players. These values are then used to calculate player strengths. Strength is then used to calculate the winning chances.

Functions used: map(), filter(), reduceByKey(), flatMap(), updateStateByKey(), union(), foreachRDD().

Algorithms used: Word count

Results

Below are the results we got:

- Player profile
- Player rating
- Chemistry
- Predicting winning chances

We saw that very few own goals and fouls were committed due to which there was less effect on player ratings.

Problems

We did not have any issues with streaming the data but the coding is very different from functional programming due to which there were many problems faced in calculations of required attributes. Time given was insufficient to come up with results. Very little information provided in the instruction ppt.

Conclusion

We learnt a lot about functional programming. For any application-based spark project, takes a good knowledge about the internal workings of spark and streaming spark.

EVALUATIONS:

SNo	Name	SRN	Contribution (Individual)
1	Prithwish Nag	PES1201800405	Player profile and ratings
2	Siri LY	PES1201801928	Player profile and ratings
3	Varshini G	PES1201801630	Player chemistry and prediction
4	D Aishwarya	PES1201801075	Player chemistry and prediction

(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.	