Phase II Report

Sasya Malireddy

Vijay Kumar Tummala

Praveen Kumar Surapaneni

**Introduction:**

We intend to develop a Web Application that will dynamically analyze tweets on IPL 2016 that is going to be held in India and generate interesting findings on this data.

**Design:**

Every line of code written in our project is in JAVA and Spark is the framework we have used for running big data analysis. Sparkcontext is the DB we have used for storing temporary data. All the graphs were developed using D3.js.

**Language used: JAVA**

**Frameworks: SPARK, D3.js**

**DB: SparkDB**

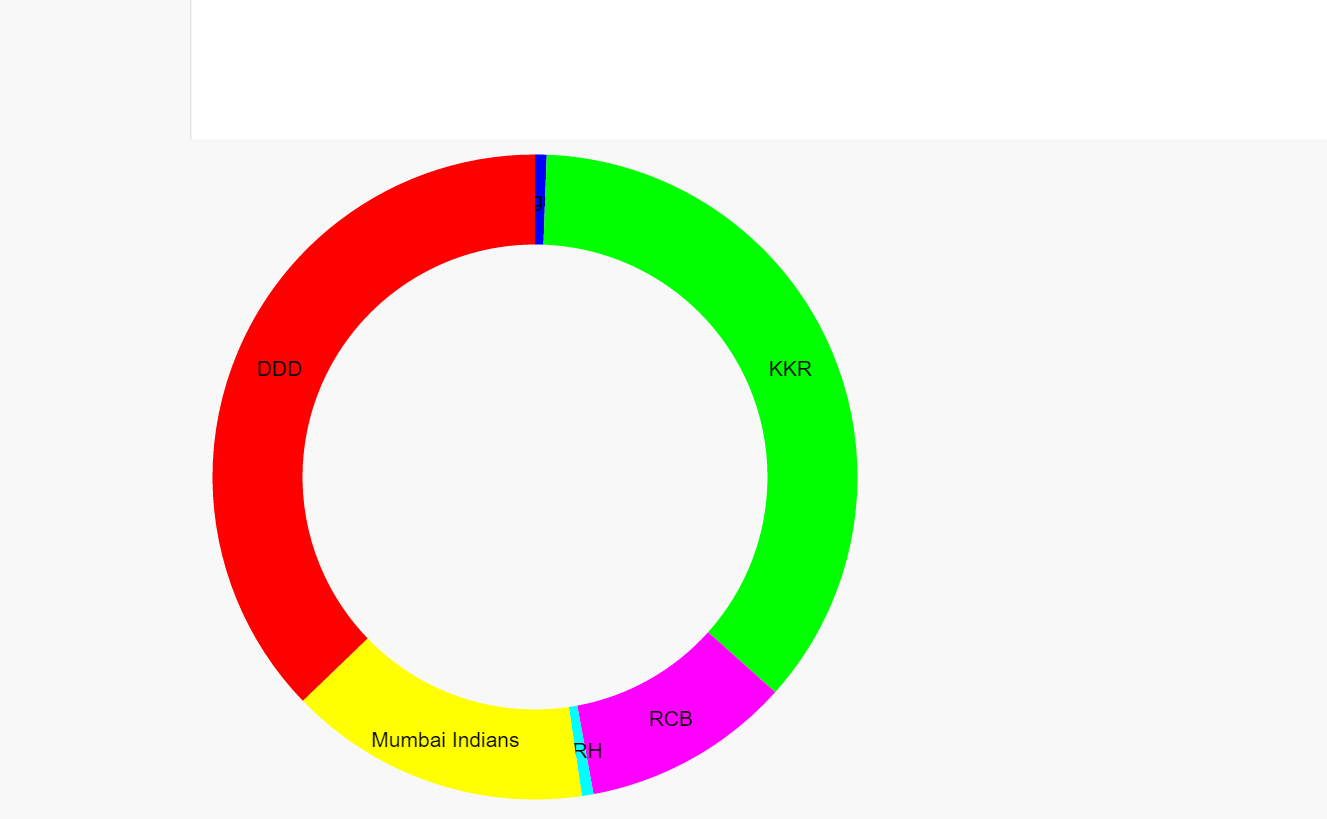
**UI: HTML, CSS**

Home page consists links for the available queries and by clicking on each link we will process an html request in the backend. For each request there will be different servlet which will handle the incoming requests. For each incoming request we will instantiate a servlet which will run a query on the spark db. Once the results are received we will create a spark context which will read the tweet data and generate the output for that particular query. We designed eight queries which will generate eight different analysis.

**Queries:**

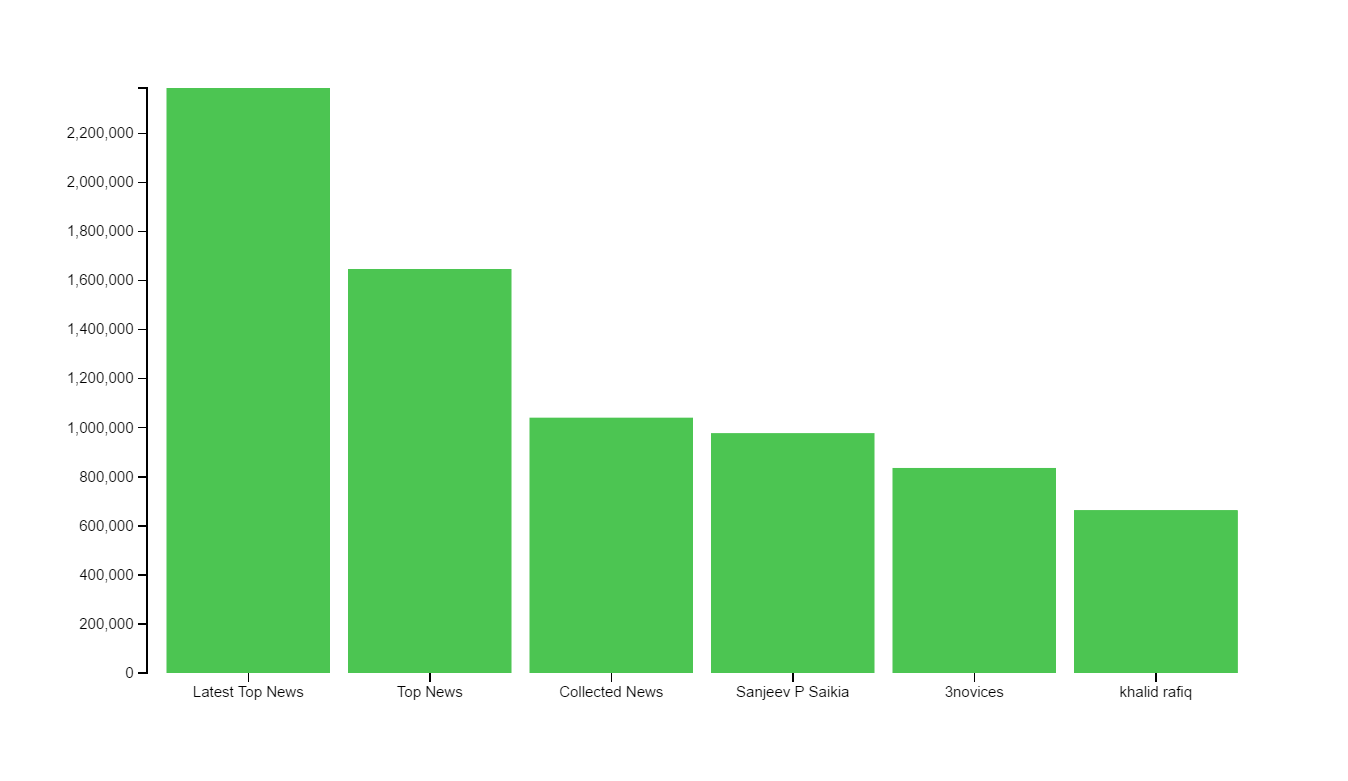
**Query 1: Most Popular IPL teams**

The popularity for the each IPL team is analyzed by collecting the count of tweets on each team. We visualize the following of all IPL teams using a Pie chart.

****

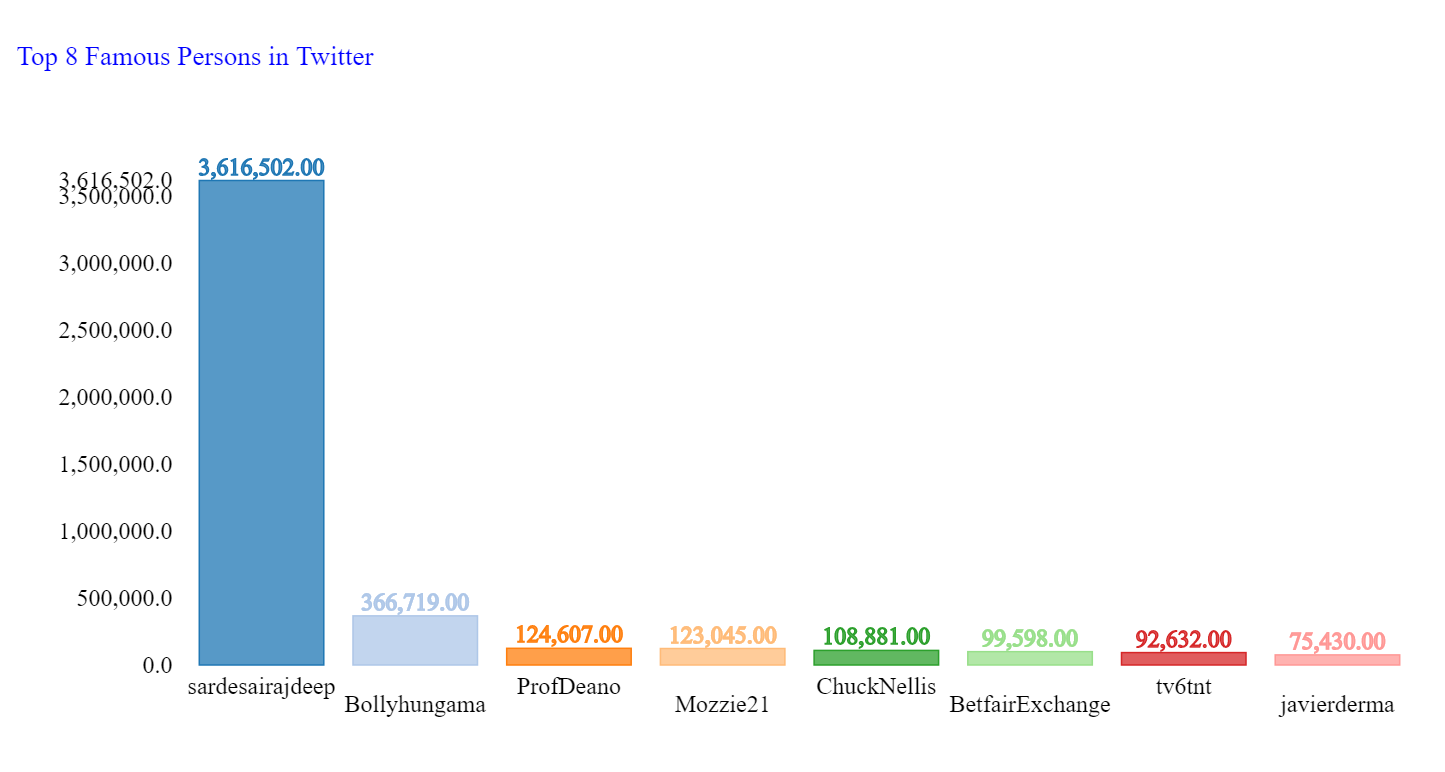
**Query 2: Top Most Users Tweeting about IPL**

The top users who tweet more about IPL on twitter are found using the following query. The Bar Graphs are used to represent these top users and their tweet counts.

****

**Query 3: Users with high follower count tweeting about IPL**

Users on twitter with highest follower count and tweeting about IPL are traced using below query. We use bar graphs for visualizing these users and there follower count statistics.



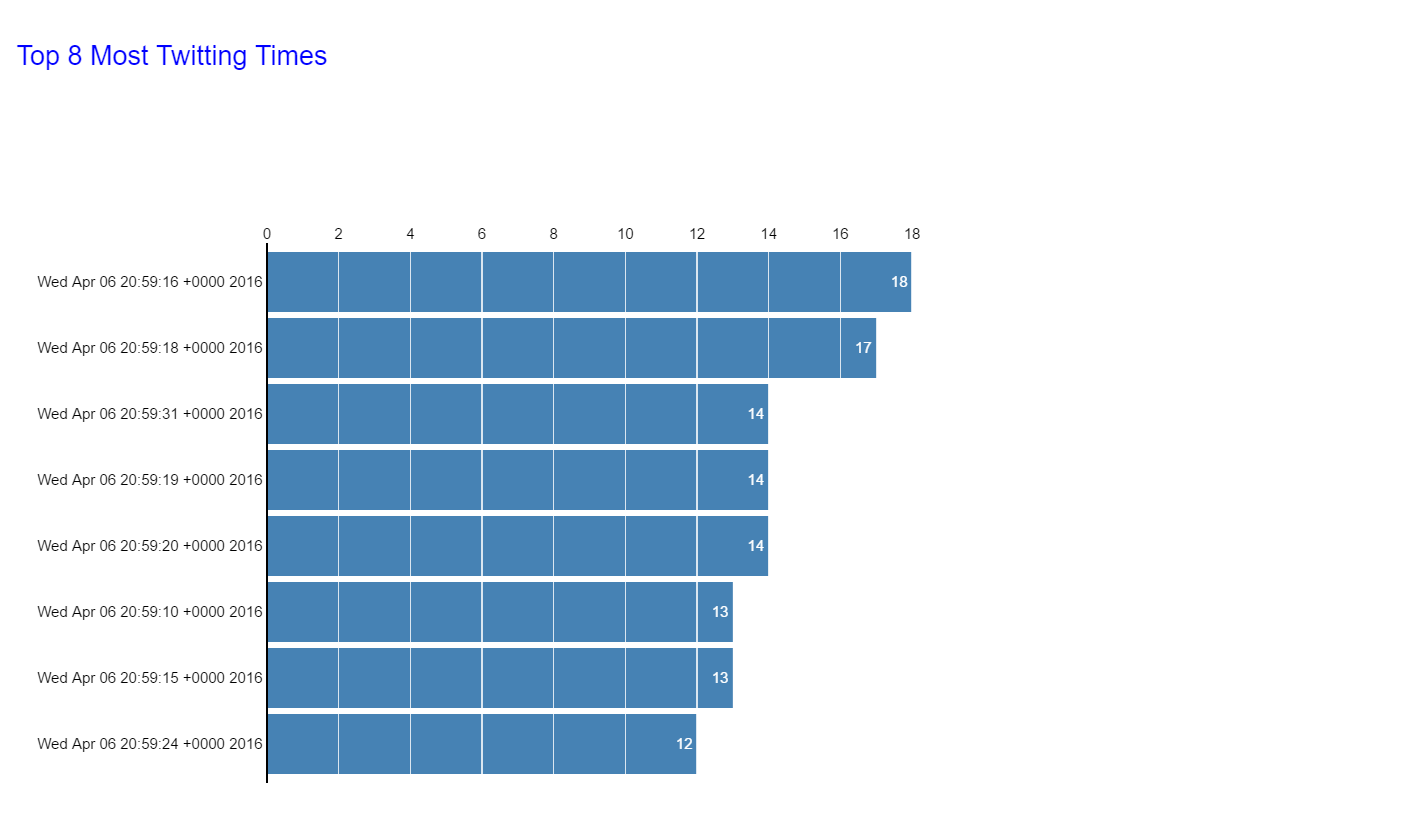
**Query 4: Positive and Negative Tweets stats about IPL on Twitter**

The comparison of positive and negative tweets is done using sentimental analysis on twitter data by some key words which exhibit the feeling. The comparative bar graph is used to visualize these stats.



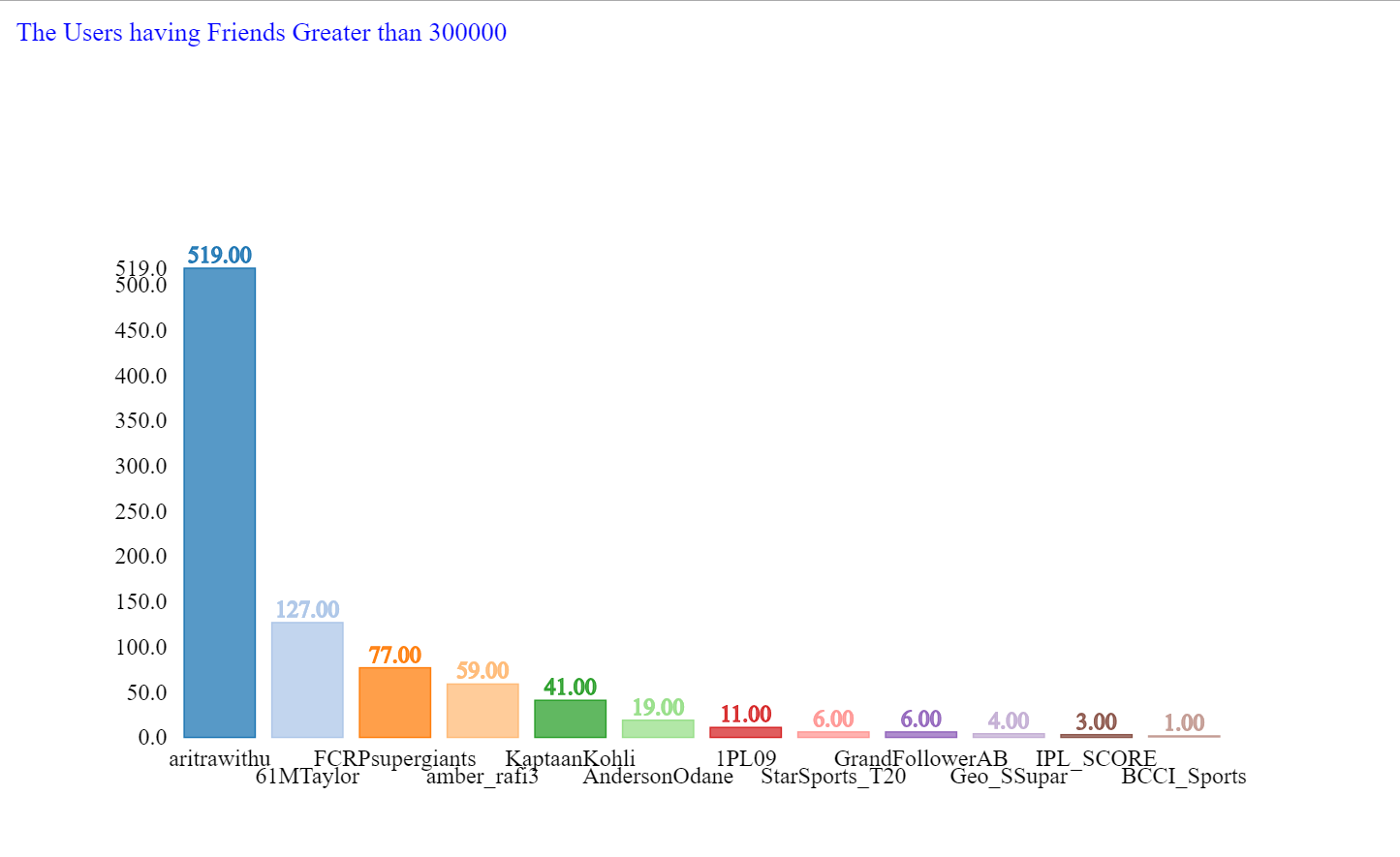
**Query 5: Most Common timings of users tweeting about IPL on twitter**

The time around clock when the users are more actively tweeting about IPL can be traced using the below query. The Horizontal bar graphs are used to visualize these stats.



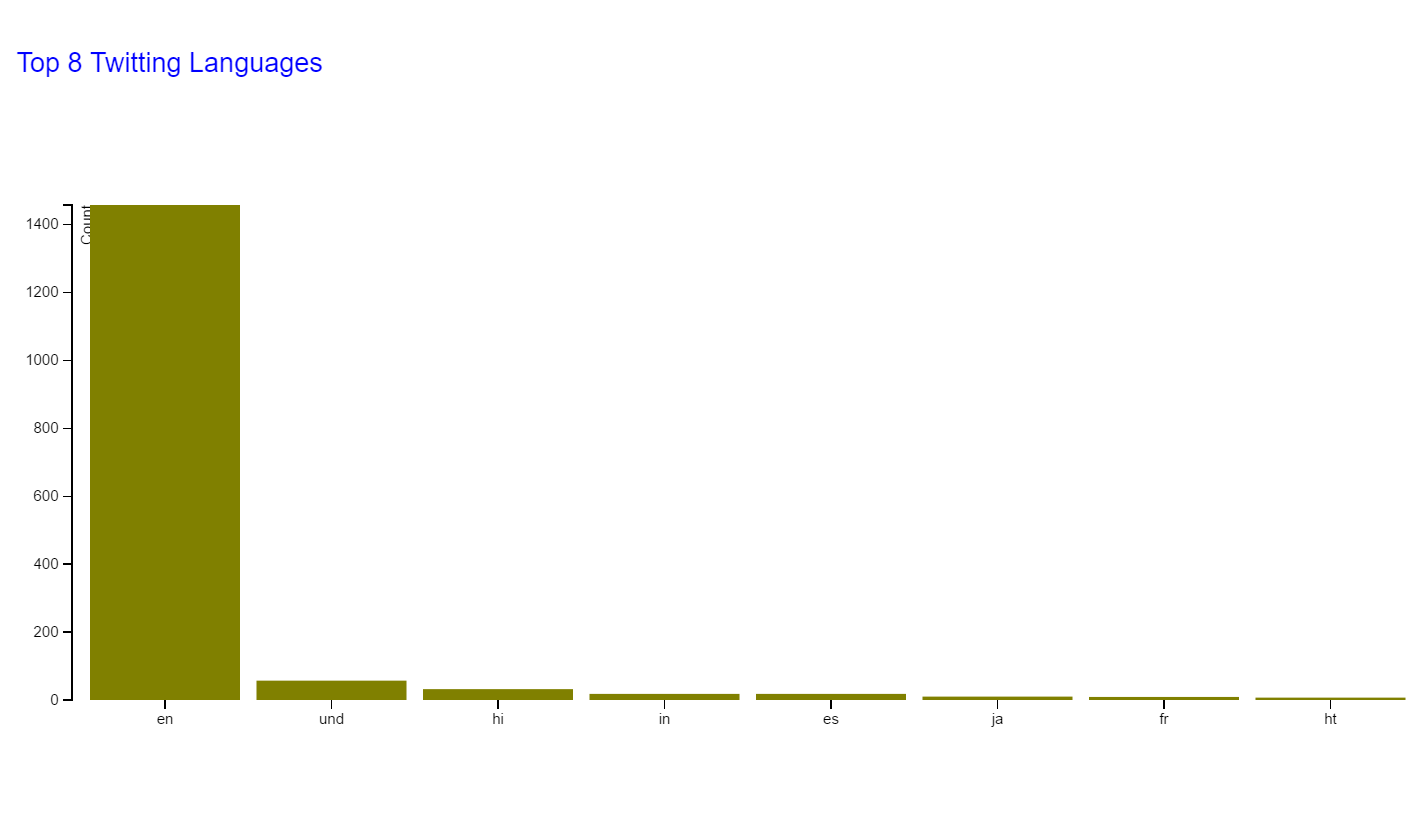
**Query 6: Users with Highest follower count supporting a single IPL Team**

Users with highest follower count who are tweeting about a particular IPL can be traced using this query. We used Bar graphs to visualize these stats.



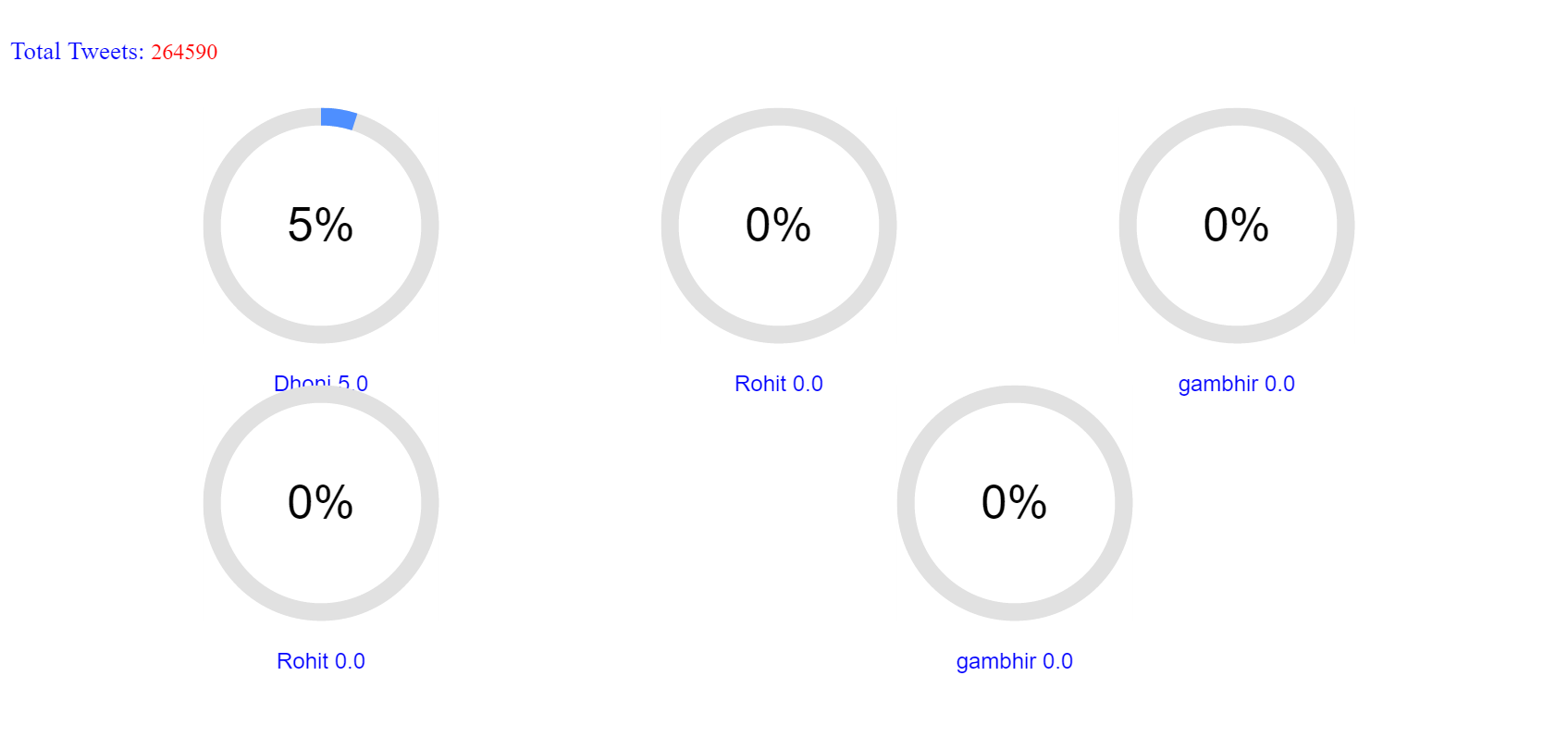
**Query 7: Language which has highest tweets about IPL**

We can trace out the languages which are high on IPL tweets using the below query. These are visualized using Bar Graphs.



**Query 8: Team Captains and there following on twitter**

The tweet count for each IPL team captain are retrieved and are compared using bar graph. These are visualized using circular graphs.

****

**References:**

<https://github.com/AshokYaganti/PBBlumixCode>

<https://github.com/stefani75/workspace/tree/b90a63f2f3028fb358b28a77ac416b223d37a52a/projet1/Hands-On-Spark-java-solution/src/main/java/com/duchessfr/spark>

**GIT Link:**

<https://github.com/Vijay-ASE/Prin_Of_BigDta/>