**Phase2 – Principles of Big Data Project Report**

**Team Members**

1. Sharath Koppu
2. Aravind Sheri
3. Harish Joshi

**Big Data** Big data is a term that describes the large volume of data – both structured and unstructured – that inundates a business on a day-to-day basis. But it’s not the amount of data that’s important. It’s what organizations do with the data that matters. Big data can be analyzed for insights that lead to better decisions and strategic business moves.

**Softwares Used**

* Apache Spark
* Scala (Execution of Queries)
* D3 JS (Data Driven Documents) for Visualizations
* High Charts for Visualizations
* HTML, CSS, Javascript for front end UI
* Webstorms for web application
* Twitter4J Library for Tweets Extraction

**Hardwares Used**

* Windows Machine
* Intel i7 Processor
* 8GB RAM

**Implementation**

1. Written a Java Program using Twitter 4J Library to extract Streaming Twitter JSON data of 5000 Tweets (20 MB) for the trending topic “#KXIPvsSRH” an Indian Premier League cricket match.
2. The extracted JSON tweets are persisted into the Apache Spark in the form of tables/views.
3. Created a web application with buttons and on the click of each button, a query written in Scala language will be sent to spark server and the outputs files are stored in the form of CSV/JSON files.
4. These CSV/JSON output files are used to visualize the data using D3 JS, Highcharts & Amcharts to get the Bar, Pie, Line, Donut Graphs etc..

**Concept – Indian Premier League (IPL2k18)**

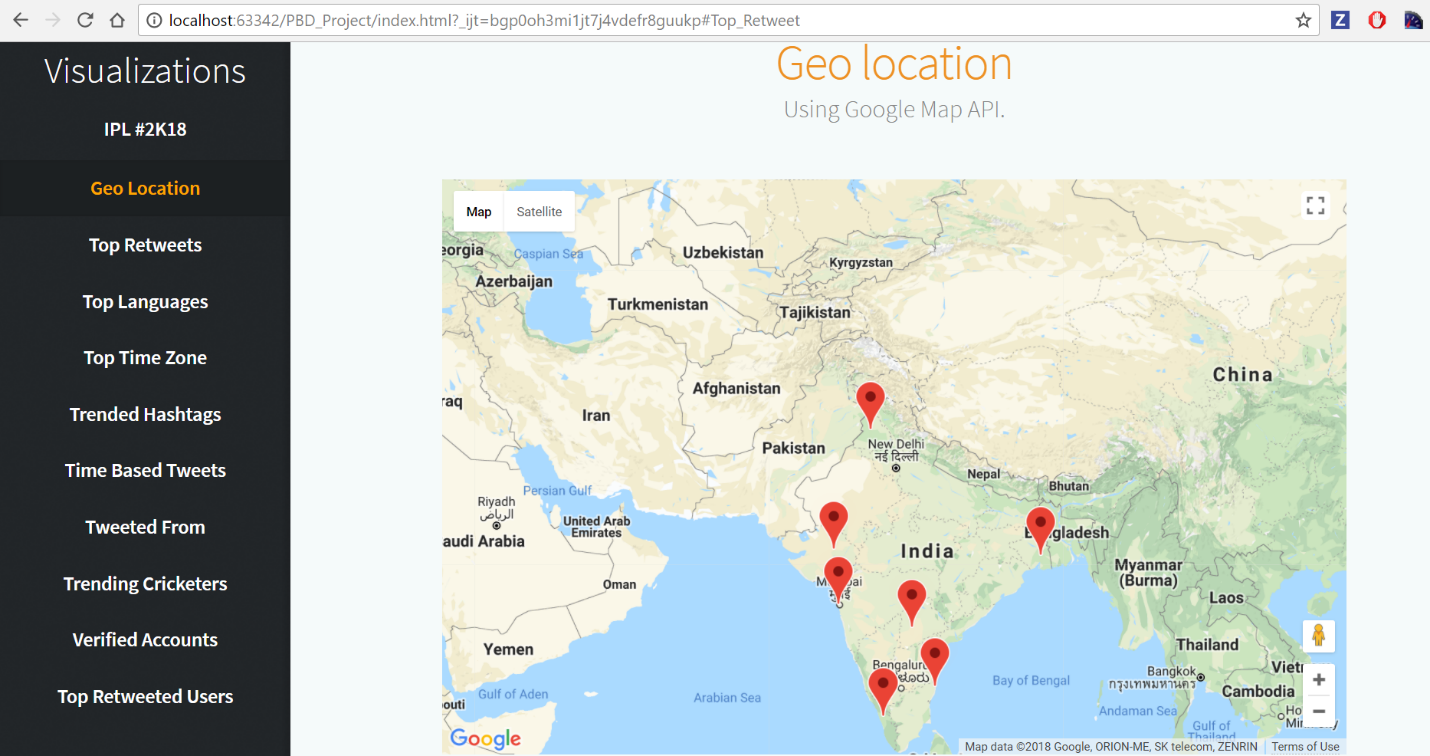
* The 2018 season of the Indian Premier League, also known as IPL 11, is the current eleventh season of the IPL, a professional Twenty20 cricket league established by the BCCI in 2007. The season, which is being held from 7 April to 27 May, saw the return of the Chennai Super Kings and the Rajasthan Royals after serving two years of suspension for the involvement of their respective owners in the 2013 IPL betting case. Star Sports purchased the media rights at ₹16,347.5 crore ($2.55 billion) for five years starting from 2018.
* We have used a trending tag “KXIPvsSRH”, the match was between Sun Risers Hyderabad & Kings eleven Punjab that was held on April 26th, 2018

**Data Analytics on IPL Cricket Match using Twitter Data**

**Query 1: IPL Trending Tweets from Top 10 Geo Locations**

val geoloc = spark.sql("SELECT user.name,user.screen\_name,user.id,geo.coordinates[0] As latt,geo.coordinates[1] As long from TwitterAnalysis where geo is not null")

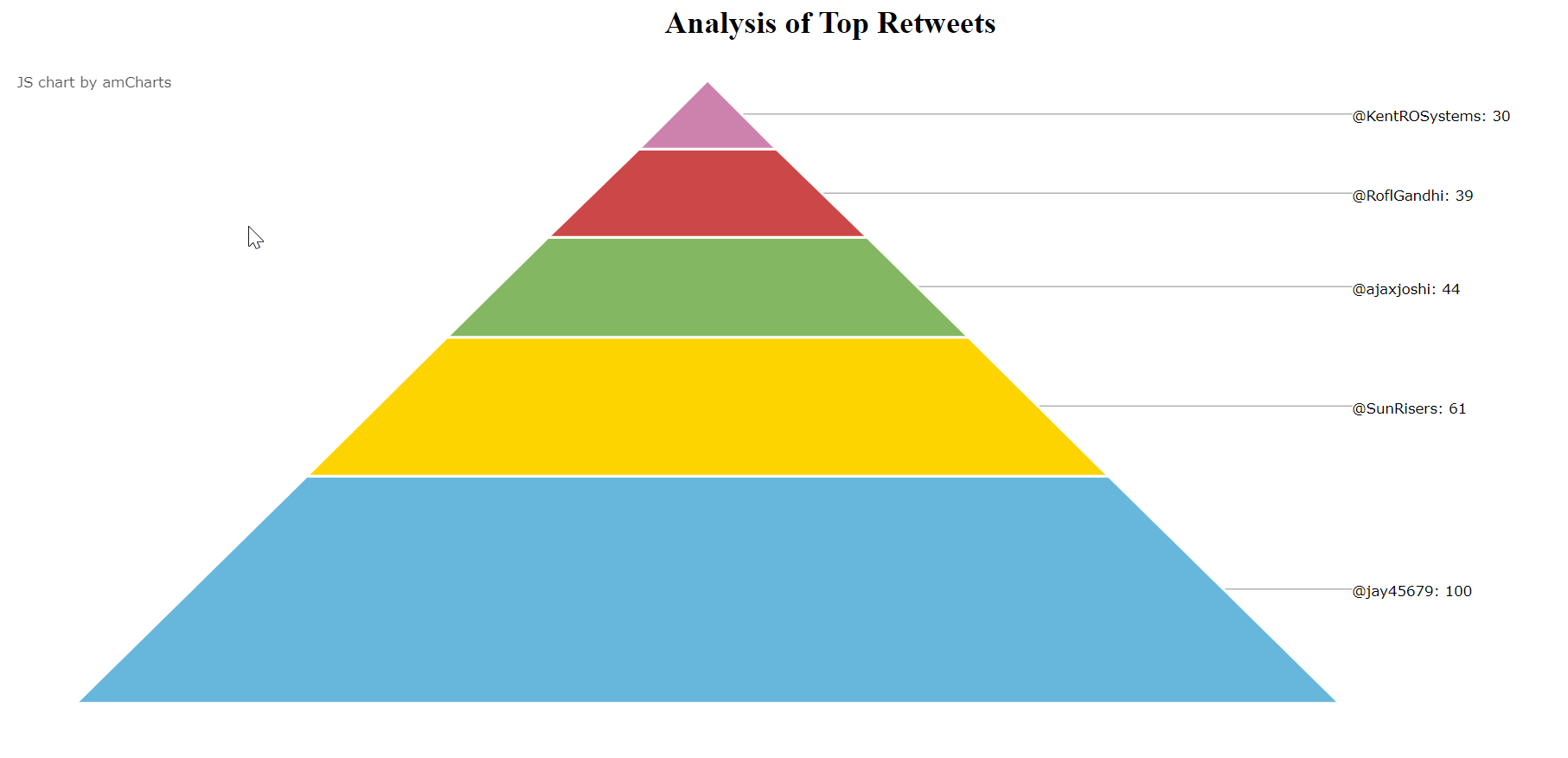
This is used to get the top 10 geo locations and show it on the Google Maps from where the trending has happened.



**Query 2: IPL Trending Top Retweets**

val retweets = spark.sql("SELECT retweeted\_status.id, retweeted\_status.text, COUNT(\*) as total FROM TwitterAnalysis WHERE retweeted\_status.user is not null GROUP BY retweeted\_status.id, retweeted\_status.text ORDER BY total desc LIMIT 5")

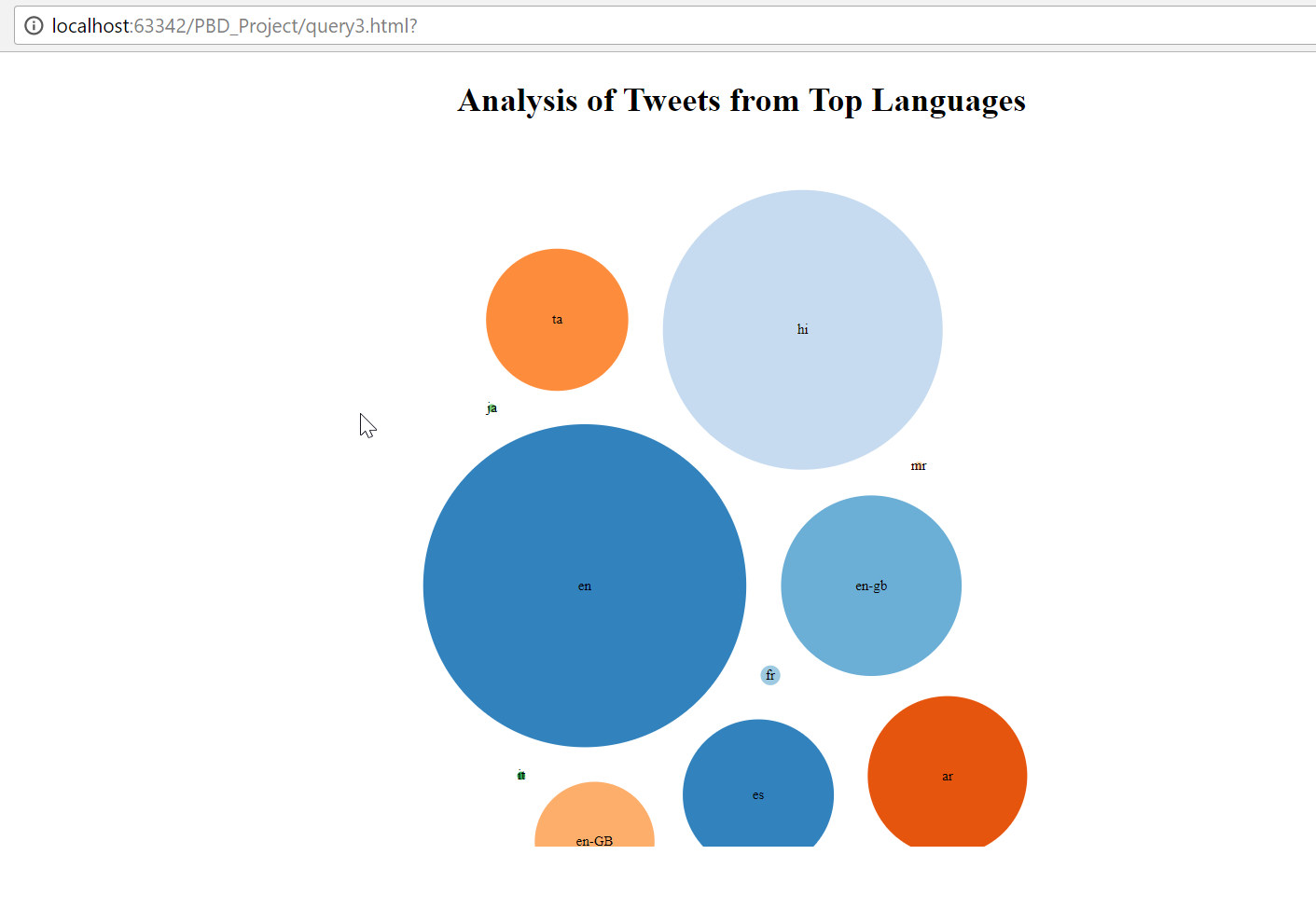
This query is used to see the top retweets and their values in the form of pyramid shape.



**Query3: IPL Trending Tweets from Different Languages**

var langs = spark.sql("SELECT user.lang, COUNT(\*) as cnt FROM TwitterAnalysis GROUP BY user.lang ORDER BY cnt DESC LIMIT 25")

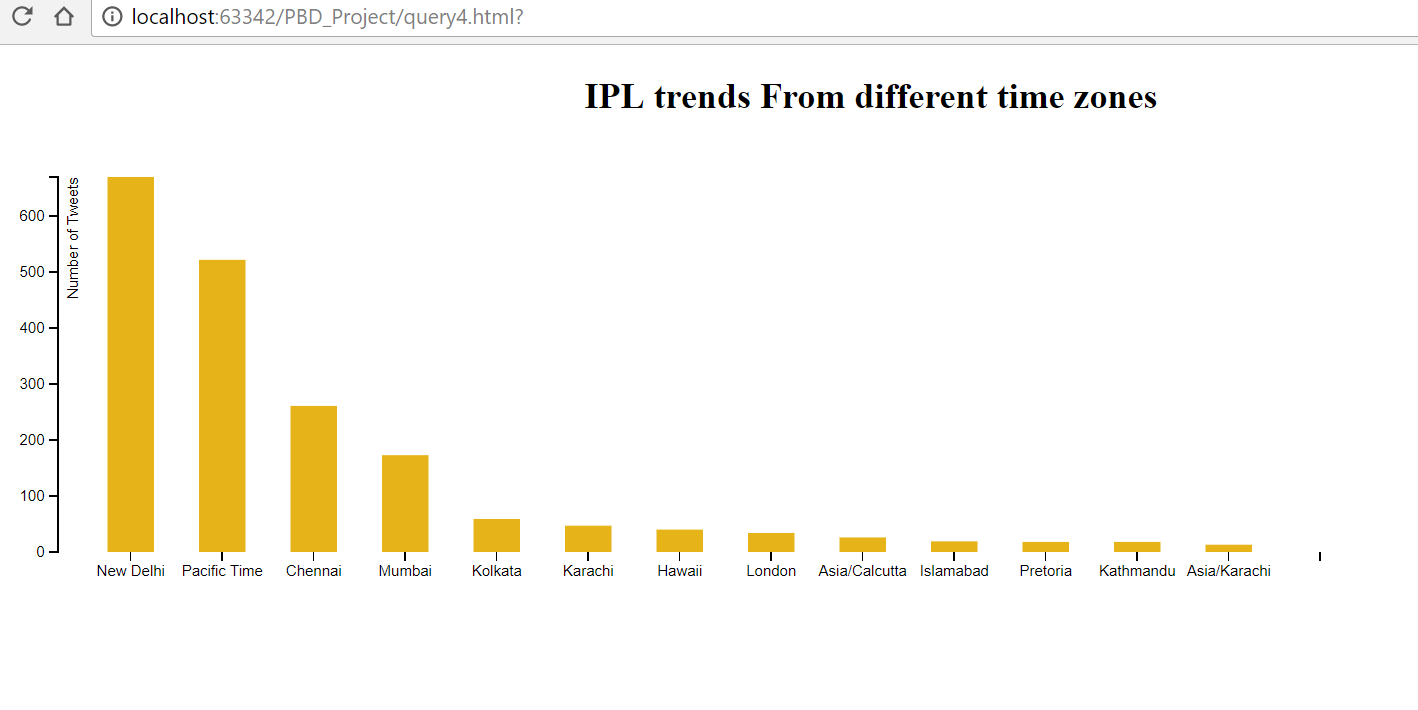
This is used to know how many tweets related to cricket have been trended from different types of languages.



**Query4: IPL Trended Tweets from Different Timezones**

val timezones = spark.sql("SELECT user.time\_zone, COUNT(\*) AS total\_count FROM TwitterAnalysis WHERE user.time\_zone IS NOT NULL GROUP BY user.time\_zone ORDER BY total\_count DESC LIMIT 50")

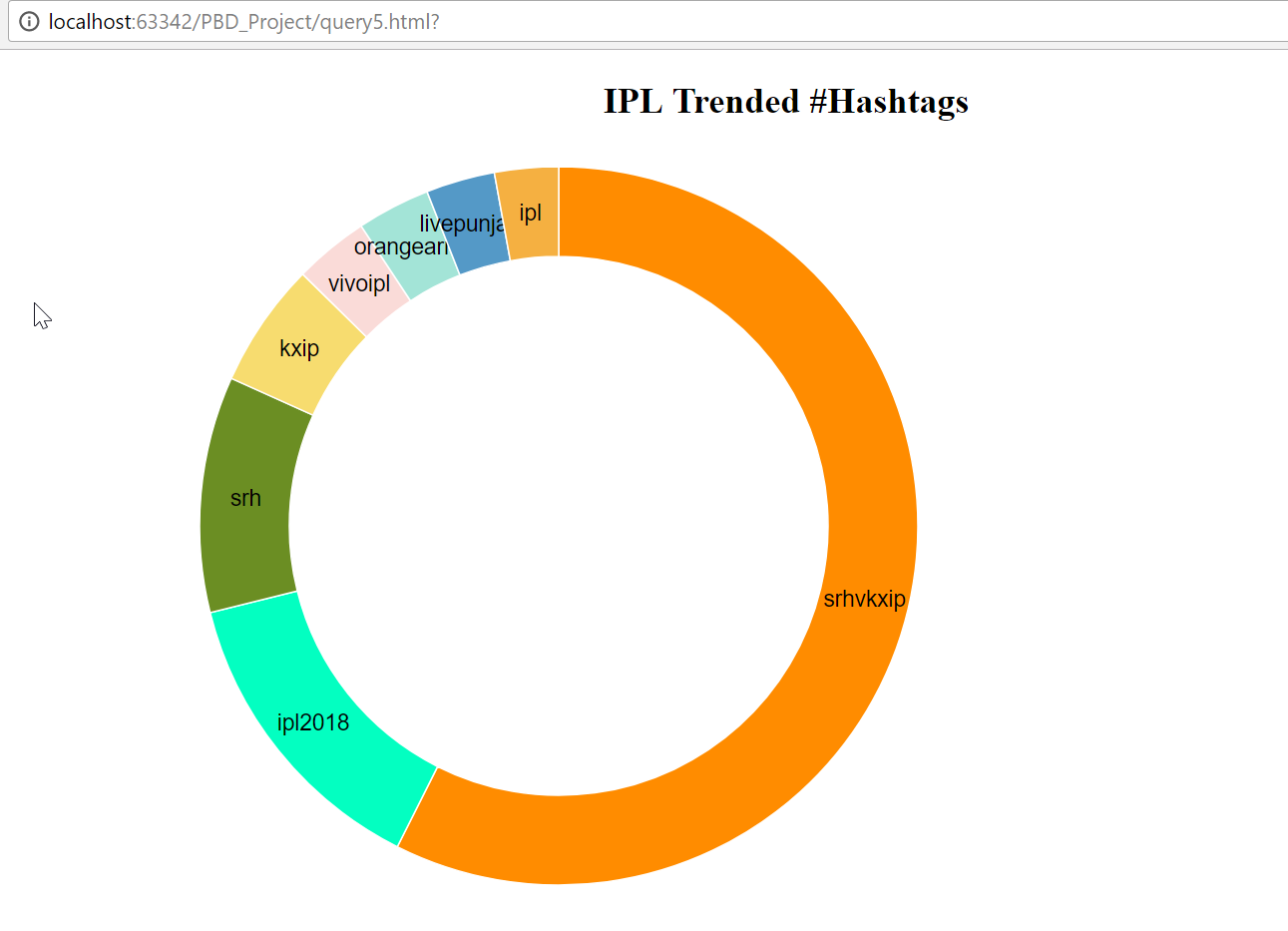
This query is used to know how many tweets have been came from different timezones of the world.



**Query5: Top Trended Cricket Match hashtags**

val hashtags = spark.sql("SELECT LOWER(hashtags.text) As Hashtags, COUNT(\*) AS total\_count FROM TwitterAnalysis LATERAL VIEW EXPLODE(entities.hashtags) AS hashtags GROUP BY LOWER(hashtags.text) ORDER BY total\_count DESC LIMIT 20")

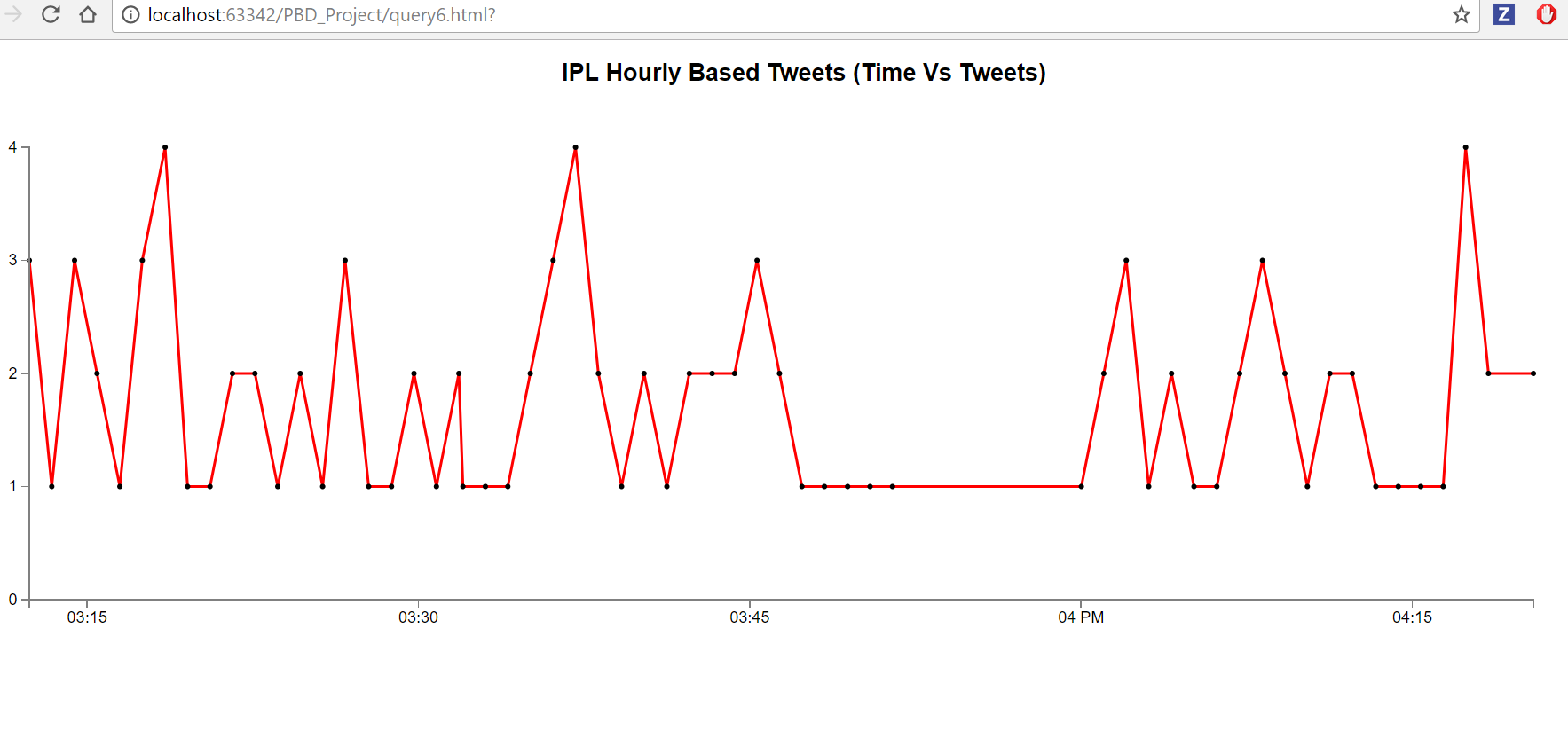
This query is used to know how many hashtags related to cricket has been trended through worldwide.



**Query6: Hourly based Tweets**

var time= spark.sql("SELECT SUM(case when created\_at >= 'Thu Apr 26 15:00%' AND created\_at<= 'Thu Apr 26 15:30%' then 1 else 0 end)Time\_Apr\_26\_1500\_1530,SUM(case when created\_at >= 'Thu Apr 26 15:30%' AND created\_at<= 'Thu Apr 26 16:00%' then 1 else 0 end)Time\_Apr\_26\_1530\_1600,SUM(case when created\_at >= 'Thu Apr 26 16:00%' AND created\_at<= 'Thu Apr 26 16:30%' then 1 else 0 end)Time\_Apr\_26\_1600\_1630,SUM(case when created\_at >= 'Thu Apr 26 16:30%' AND created\_at<= 'Thu Apr 26 17:00%' then 1 else 0 end)Time\_Apr\_26\_1630\_1700 from TwitterAnalysis")

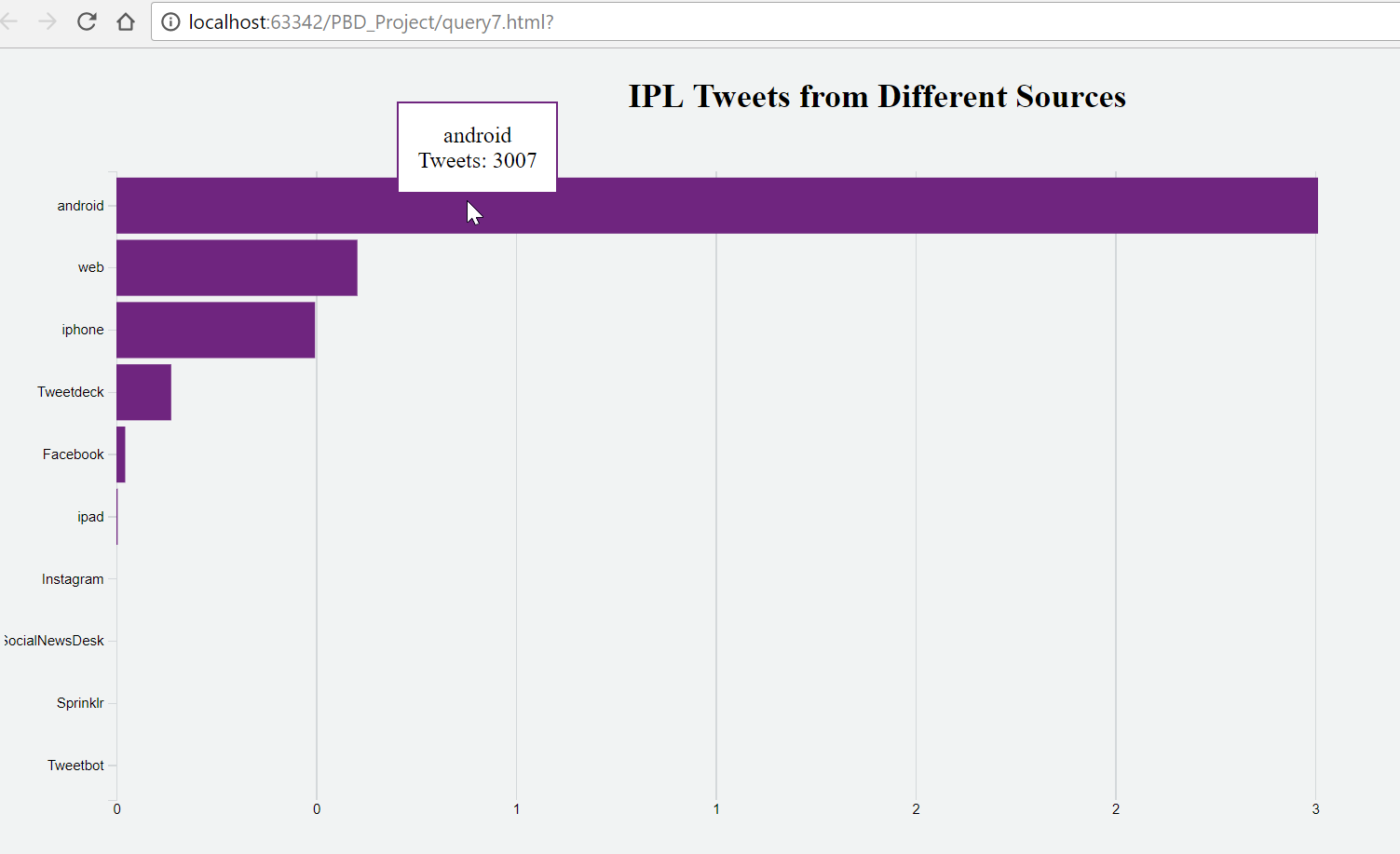
This query is used to know how many tweets have been generated for a timestamp. We have showed the top trends for each time division.



**Query7: Analysis on Twitter Sources**

val sources = spark.sql("SELECT sum(case when lower(source) like '%android%' then 1 else 0 end)android, sum(case when lower(source) like '%iphone%' then 1 else 0 end)iphone, sum(case when lower(source) like '%ipad%' then 1 else 0 end)ipad, sum(case when lower(source) like '%web%' then 1 else 0 end)web, sum(case when lower(source) like '%tweetdeck%' then 1 else 0 end)Tweetdeck, sum(case when lower(source) like '%tweetbot%' then 1 else 0 end)Tweetbot, sum(case when lower(source) like '%www.sprinklr.com%' then 1 else 0 end)Sprinklr, sum(case when lower(source) like '%www.socialnewsdesk.com%' then 1 else 0 end)SocialNewsDesk, sum(case when lower(source) like '%instagram.com%' then 1 else 0 end)Instagram, sum(case when lower(source) like '%facebook.com%' then 1 else 0 end)Facebook from TwitterAnalysis")

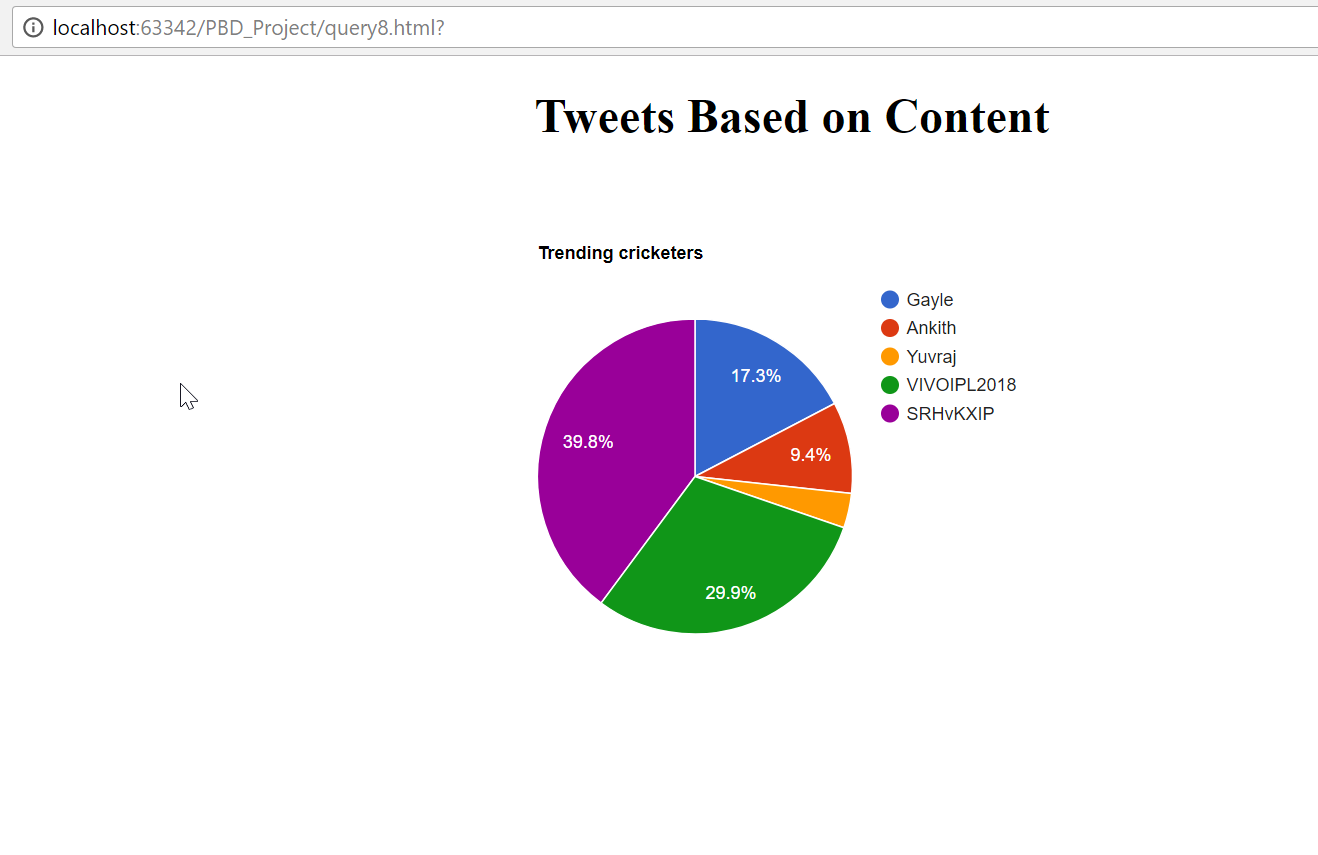
This query is used to know the number of tweets that have been generated from different platforms like ios, android, web, etc…



**Query8: Analysis on Cricket Players**

var players = spark.sql("SELECT SUM(case when text like '%Gayle%' then 1 else 0 end)Gayle,SUM(case when text like '%Ankit Rajpoot%' then 1 else 0 end)Ankith,SUM(case when text like '%Yuvraj%' then 1 else 0 end)Yuvraj,SUM(case when text like '%VIVOIPL2018%' then 1 else 0 end)VIVOIPL2018,SUM(case when text like '%SRHvKXIP%' then 1 else 0 end)SRHvKXIP from TwitterAnalysis")

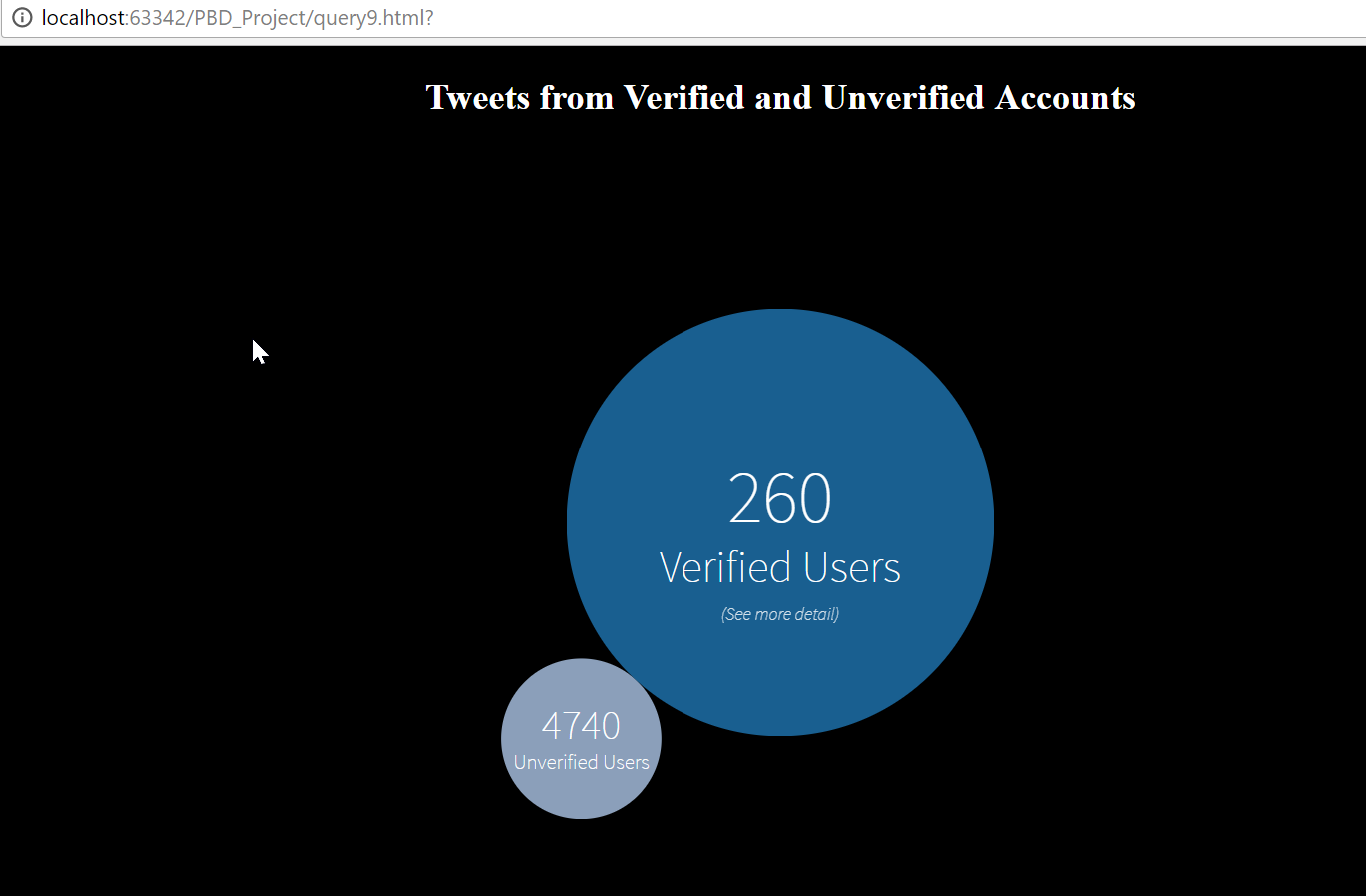
This query is used to know the top cricket players that has been trended through out the match.



**Query9: IPL Tweets from verified Users**

var status = spark.sql("SELECT SUM(case when user.verified=true then 1 else 0 end)Verified,SUM(case when user.verified=false then 1 else 0 end)Unverified from TwitterAnalysis")

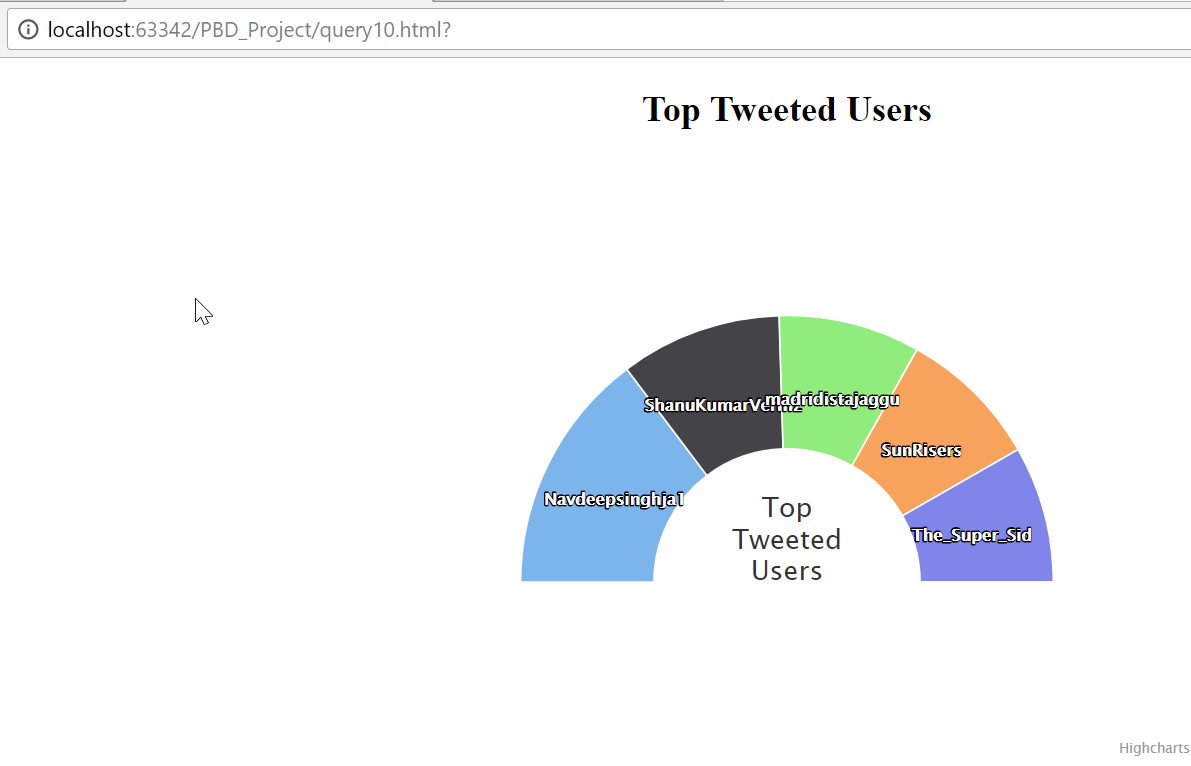
This query is used to know the tweets trending from verified/unverified users



**Query10: IPL Top Tweeted Screens**

var screens = spark.sql("SELECT user.screen\_name, COUNT(user.screen\_name) as cnt FROM TwitterAnalysis GROUP BY text ORDER BY cnt DESC LIMIT 5")

This query is used to know how many users have tweeted the tweets the most



**Testing the Code**

1. Extract the zip file attached and you can find the Project report and the Project code folder.
2. Open the Project folder using Webstorms and click the index.html file
3. Now we can see the web application will be opened in browser and we can click on each button to see the visualizations of ten queries.