**Phase 2 Report**

**Sireesha Keesara-16277715**

**Ahamed Raziuddin**

**Bhavaz Akula**

**Abstract:**

1. Writing analytical queries on the twitter data that we collected.
2. Developing interesting visualizations.

**Technologies used:**

1. Spark
2. Python
3. Tableau
4. Zeppelin

**Queries:**

**Query 1:** Number of users tweeting about cricket, football and tennis.

sqlDF = spark.sql("SELECT 'Cricket' as sport, count(\*) as Count from Sports where text like '%cricket%' or text like '%ipl%' or upper(text) like '%CRICKET%' or upper(text) like '%IPL'\

UNION\

SELECT 'Tennis' as sport, count(\*) as Count from Sports where text like '%tennis%' or upper(text) like '%TENNIS%'\

UNION\

SELECT 'Football' as sport, count(\*) as Count from Sports where text like '%football%' or upper(text) like '%FOOTBALL%' or text like '%fifa%' or upper(text) like '%FIFA%'")

pd = sqlDF.toPandas()

A screenshot of a cell phone

Description automatically generated

**Query2:** Number of tweets from top 10 users

sqldf = spark.sql(

"SELECT user.id,user.name,count(\*) FROM Users"

" WHERE (user.id is not null and user.name is not null) group by user.id,user.name order by 3 desc limit 10")

sqldf.show(150)

A close up of a sign

Description automatically generated

**Query3:** Number of tweets on Players like Virat, MSD, Russell, ABD, Sachin Tendulakar and Shikar Dhawan.

sqlhash = spark.sql("SELECT 'Virat' player,count(text) as count \

FROM cricketplayer\_count\

WHERE 1=1\

AND (upper(text) LIKE '%KHOLI%' or upper(text) LIKE '%VIRAT%' or upper(text) LIKE '%VIR%')\

GROUP BY player\

UNION\

SELECT 'Dhoni' player,count(text) as count \

FROM cricketplayer\_count\

WHERE 1=1\

AND (upper(text) LIKE '%MSD%' or upper(text) LIKE '%DHONI%')\

GROUP BY player\

UNION\

SELECT 'Russell' player,count(text) as count \

FROM cricketplayer\_count\

WHERE 1=1\

AND (upper(text) LIKE '%RUSSELL%' or text LIKE '%russell%')\

GROUP BY player\

UNION\

SELECT 'Sachin' player,count(text) as count\

FROM cricketplayer\_count\

WHERE 1=1\

AND (upper(text) LIKE '%SACHIN%' or upper(text) LIKE '%TENDULKAR%')\

GROUP BY player\

UNION\

SELECT 'ABD' player,count(text) as count \

FROM cricketplayer\_count\

WHERE 1=1\

AND (upper(text) LIKE '%ABD%' or text LIKE '%abd%')\

GROUP BY player\

UNION\

SELECT 'Shikar' player,count(text) as count \

FROM cricketplayer\_count\

WHERE 1=1\

AND (upper(text) LIKE '%SHIKAR%DAWAN%' or upper(text) LIKE '%SHIKAR%' or upper(text) LIKE '%DHAWAN%')\

GROUP BY player")

sqlhash.show()

A screenshot of a cell phone

Description automatically generated

**Query 4:**  To display the number of tweets on different formats of cricket like ODI, IPL and Test Match.

sqldf= spark.sql("SELECT 'ODI' team,count(\*) FROM cricket WHERE upper(text) LIKE '%ODI%' \

UNION \

SELECT 'TEST' team,count(\*) FROM cricket WHERE upper(text) LIKE '%TEST%MATCH%' \

UNION \

SELECT 'T20' team,count(\*) FROM cricket WHERE upper(text) LIKE '%T20%'")

sqldf.show(150)

**A close up of a logo

Description automatically generated**

**Query5:** To display the number tweets on football players Messi and Ronaldo.

sqldf = spark.sql("SELECT 'MESSI' player,count(text) as count \

FROM football\

WHERE 1=1\

AND (upper(text) LIKE '%LIONEL%' or upper(text) LIKE '%MESSI%' or text like '%messi%')\

GROUP BY player\

UNION\

SELECT 'Ronaldo' player,count(text) as count \

FROM football\

WHERE 1=1\

AND (upper(text) LIKE '%RONALDO%' or upper(text) LIKE '%CRISTIANO%' or text like '%ronaldo%')\

GROUP BY player")

**A screenshot of a cell phone

Description automatically generated**

**Query6:** To display the number of tweets in different languages.

sqldf= spark.sql("SELECT user.lang language,count(\*) no\_of\_tweets FROM sports WHERE user.lang is NOT NULL GROUP BY user.lang ORDER BY 2 DESC limit 10")

**A screenshot of a cell phone

Description automatically generated**

**Query7:** To display the number of retweets of top ten users.

sqldf = spark.sql(

"SELECT name,SUM(cnt) as retweet FROM (SELECT quoted\_status.user.screen\_name AS name,quoted\_status.retweet\_count AS cnt FROM sports WHERE quoted\_status.retweet\_count>0)GROUP BY name ORDER BY retweet DESC LIMIT 10")

**A close up of text on a white background

Description automatically generated**

**Query8:** To display the number of tweets on dhoni and cricket, messi and football, serena and tennis.

sqlDF = spark.sql("SELECT 'Cricket' as Domain, count(\*) as Count from Sports where text like '%dhoni%' and text like '%cricket%'\

UNION\

SELECT 'Football' as Domain, count(\*) as Count from Sports where text like '%Messi%' and text like '%football%'\

UNION\

SELECT 'Tennis' as Domain, count(\*) as Count from Sports where text like '%serena%' and text like '%tennis%'")

**A close up of a logo

Description automatically generated**

**Query9:**  Collected the tweets on India and Cricket and done sentiment analysis on the collected data to know the people who are supporting india.

sqldf= spark.sql("SELECT id,text,created\_date FROM cricket WHERE 1=1 AND (upper(text) LIKE '%INDIA%'AND text LIKE '%cricket%')")

for t in sqldf.select("text").collect():

i=i+1

# print("It is ",i,str(t.text))

analysis = TextBlob(str((t.text).encode('ascii', 'ignore')))

print(analysis.sentiment.polarity)

if (analysis.sentiment.polarity<0):

negative=negative+1

print(i," in negative")

elif(analysis.sentiment.polarity==0.0):

neutral=neutral+1

print(i," in neutral")

elif(analysis.sentiment.polarity>0):

positive=positive+1

print(i," in positive")

print("Total negative % is",((negative)\*100)/i)

print("Total neutral % is",((neutral)\*100)/i)

print("Total positive % is",((positive)\*100)/i)

negative\_percent=((negative)\*100)/i

positive\_percent=((positive)\*100)/i

neutral\_percent=((neutral)\*100)/i

A screenshot of a cell phone

Description automatically generated

**Query10:** To display the most tweeted top 10 hashtags count.

sqlhash = spark.sql("SELECT tag\_value,\

times\

FROM (SELECT upper(tags) tag\_value,\

count(\*) times\

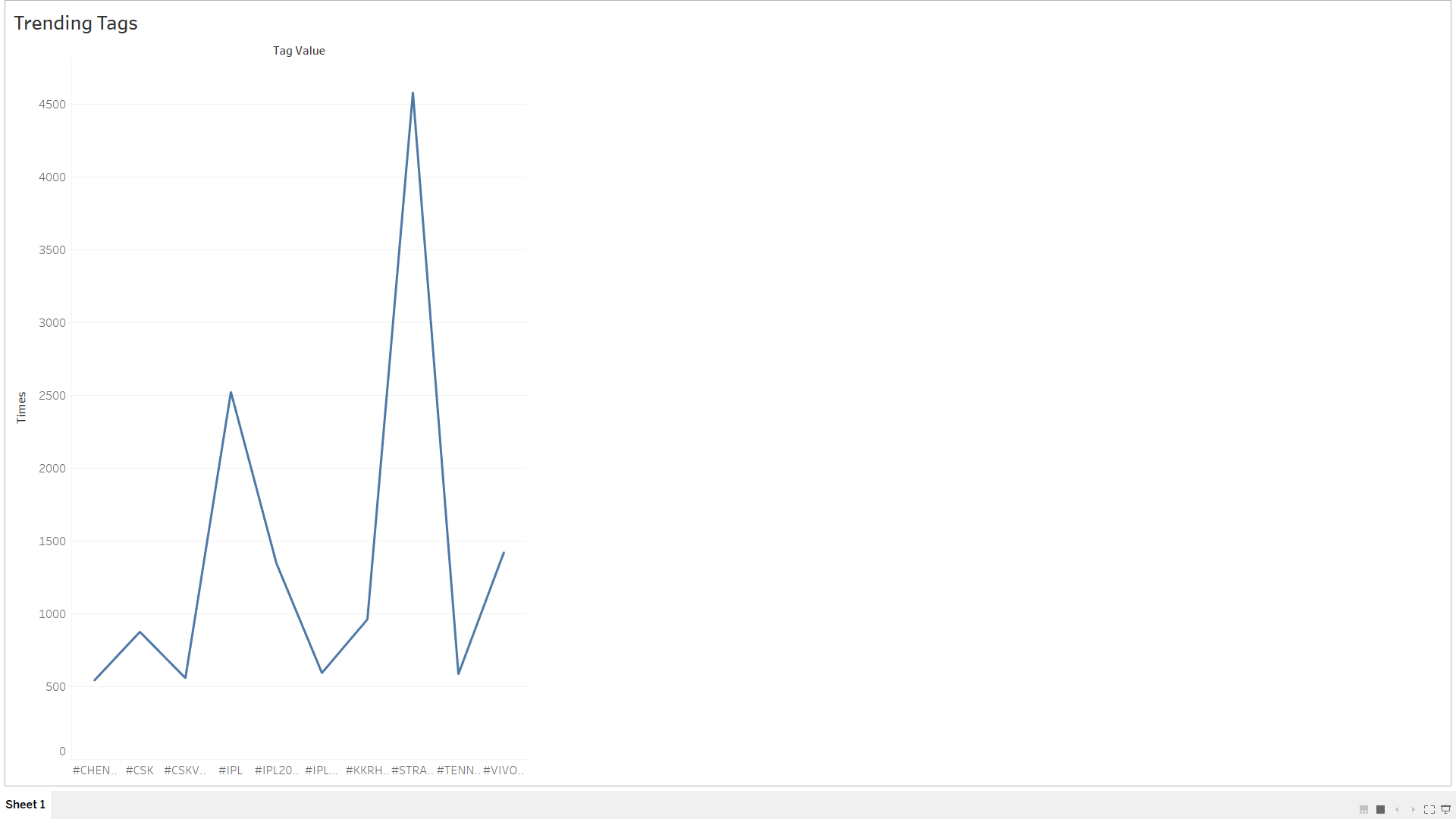
FROM hashtag\_count\

WHERE 1=1\

AND tags!='nonTag'\

GROUP BY upper(tags)\

ORDER BY times desc, tag\_value asc) limit 10")

****

**Query 11:** To display the Number of tweets from different locations.

select

case when country='India' then 'in'

when country='United States' then 'us'

when country='United Kingdom' then 'gb'

when country='Brasil' then 'br'

when country='South Africa' then 'za'

when country='Australia' then 'au'

when country='Italia' then 'it'

when country='France' then 'fr'

when country='Nigeria' then 'ne'

when country='Canada' then 'ca'

when country='Russia' then 'ru'

when country='China' then 'cn'

when country='Greenland ' then 'gl'

else country end as Country,case  when country='Russia' then 0

when country='China' then 0

when country='Greenland ' then 0 else count(\*) end as cnt from(select coalesce(place.country,"") country from table) where  nullif(country,"") is not null group by country order by cnt desc

**A close up of a map

Description automatically generated**