**CS 6350 Big Data Management & Analytics**

**Assignment 3**

**By:**

**SHARIQ ALI(sxa190016)  
SRIPRASATH GUJULUVA PARTHASARATHY(sxg180154)**

**Part 1 - Spark Streaming with Twitter and Kafka**

# Software Requirements

1. java version 1.8.0\_261
2. scala 2.11.8
3. Spark 2.4.0 built for Hadoop 2.7.3
4. kafka\_2.12-2.3.0
5. elasticsearch-7.4.2
6. kibana-7.4.2-darwin-x86\_64
7. logstash-7.4.2
8. stanford-corenlp\_3.5.2

# How to Execute?

Please find execution steps in README.txt

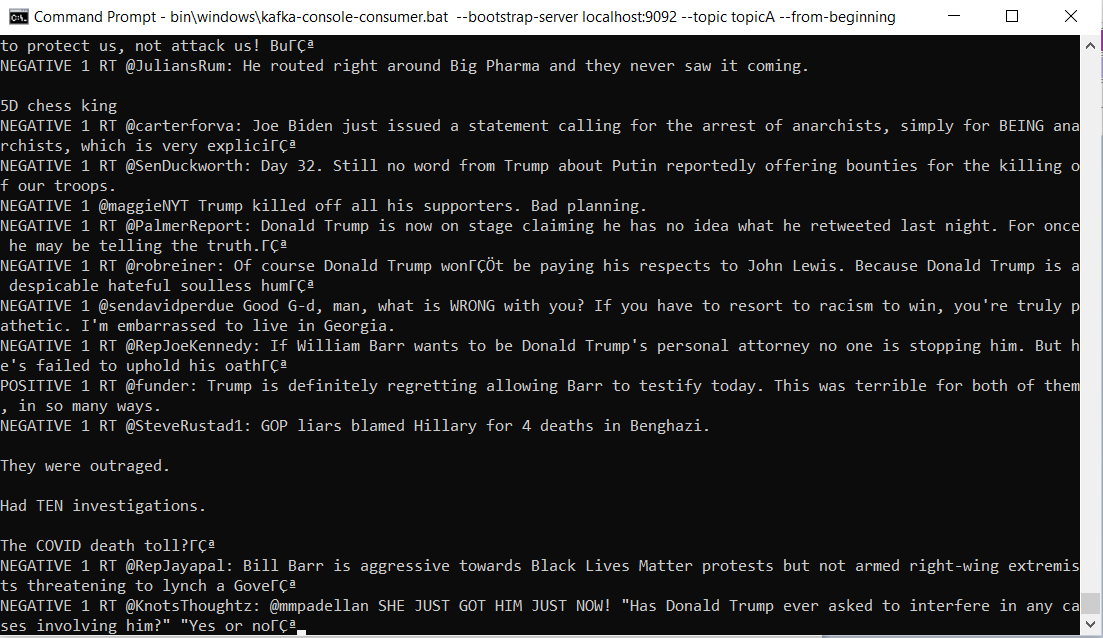
# Keywords

The tweets containing the below keywords are read streamed from Twitter: **Trump** and **Obama**

# Sequence of steps

The tweets containing the keywords Trump and Obama are streamed from Twitter. Then these tweets are analyzed using sentiment analysis to find if the sentiment of the respective tweet in positive, negative, or neutral. Then the sentiment, count and tweet are written to kafka. Then we use logstash to pipeline the records from kafka to elasticsearch. From elasticsearch we can fetch data using REST api, in JSON format, for visualization and analysis.

# Producer record format



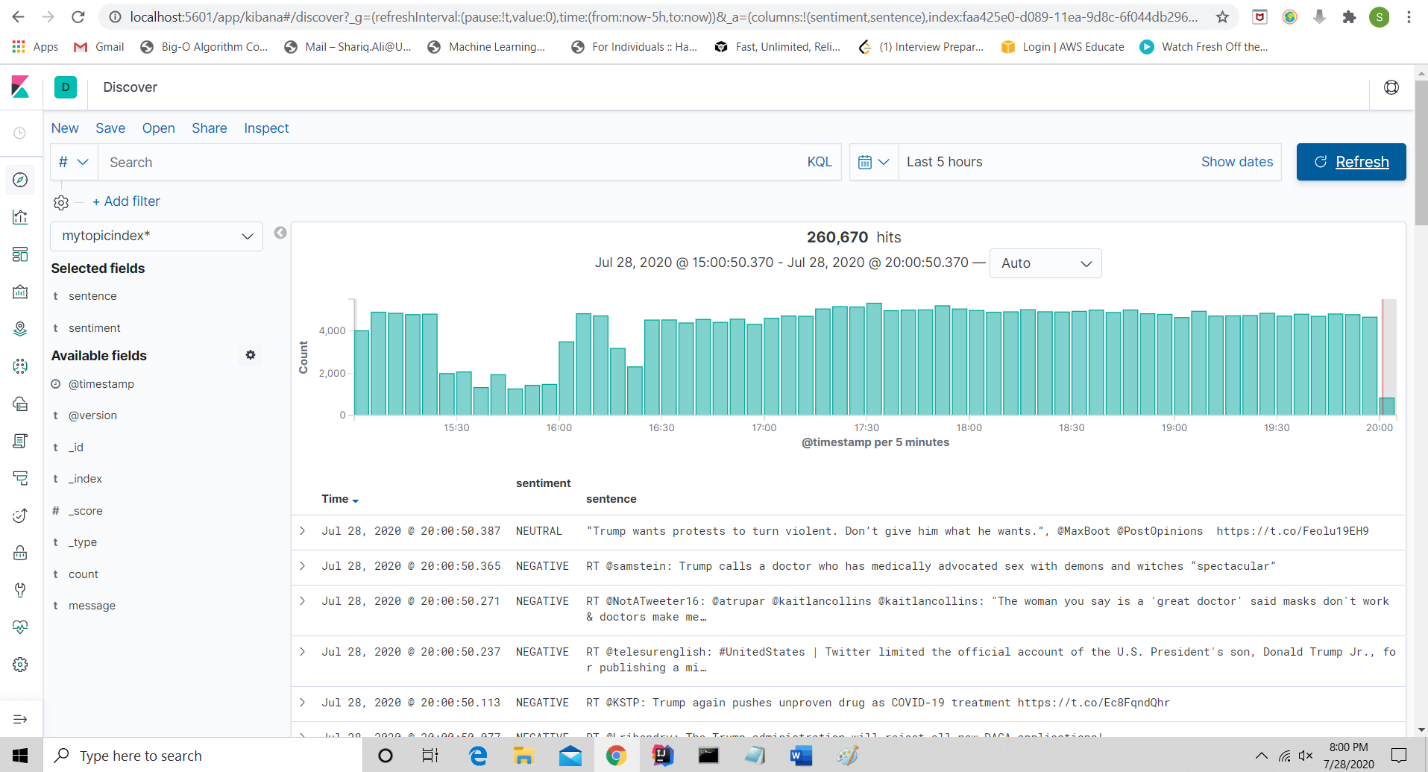
**The above is a screenshot of Producer record format as read by a consumer**

The format of the record is SENTIMENT COUNT SENTENCE (specified in logstash config file)

# Visualization in Kibana and Analysis

NOTE: To search for the data in Kibana you must create an index. The index that I have used is "**mytopicindex**".

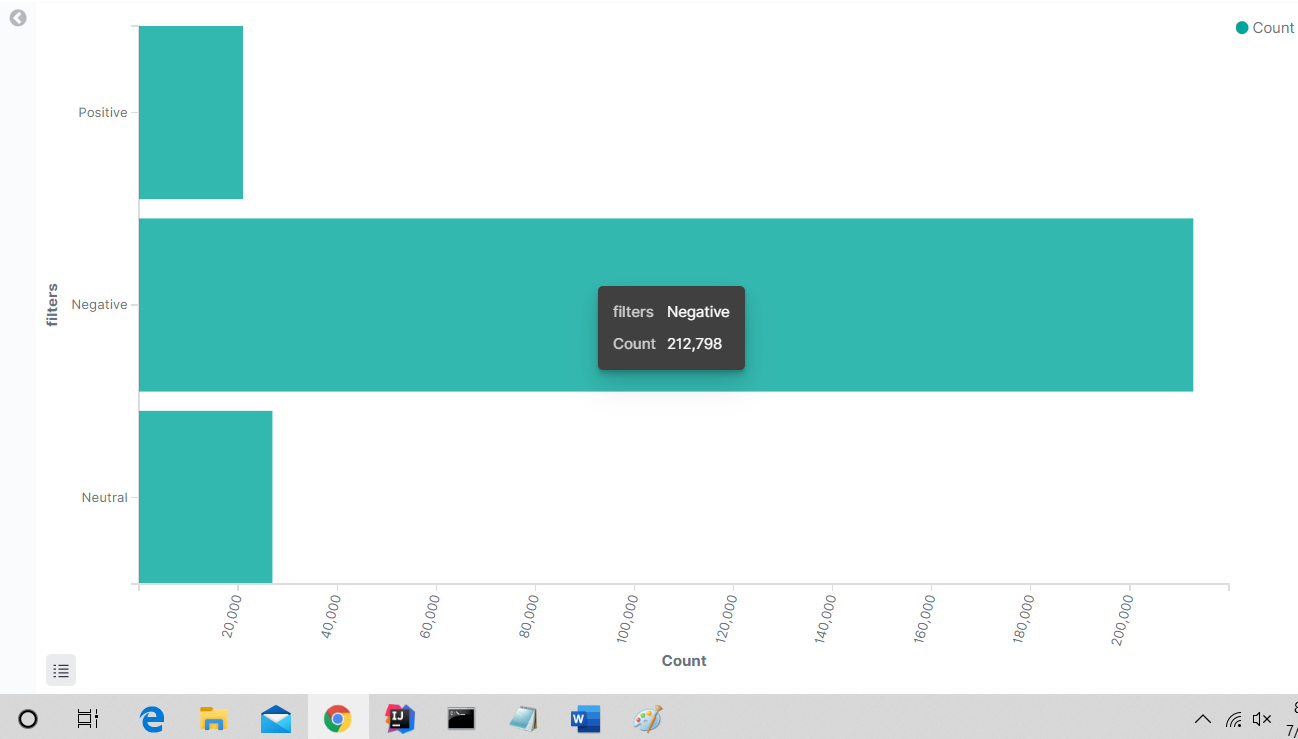
On the discover page the data looks like in the image given below:



The above screenshot shows data for the last 5 hours.

**Important metrics:**

1. Hits : 260,670 (total number of records fetched)
2. Capture time : 3:00 PM to 8:00 PM on July 28, 2020

****

The above screenshot shows **Positive Negative Neutral absolute values in last 5 hours**

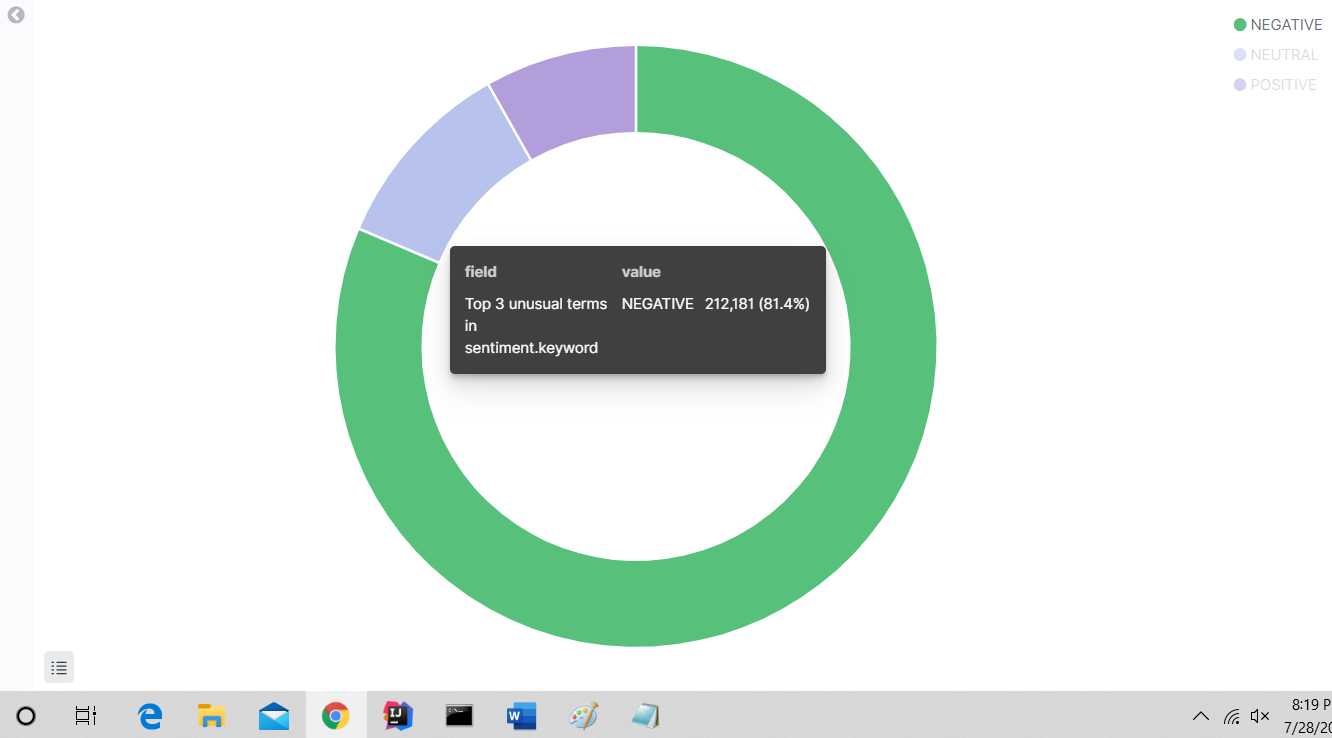
URI to recreate the visualization:

<http://localhost:5601/app/kibana#/visualize/create?type=horizontal_bar&indexPattern=faa425e0-d089-11ea-9d8c-6f044db2969a&_g=(filters:!(),refreshInterval:(pause:!t,value:0),time:(from:now-5h,to:now))&_a=(filters:!(),linked:!f,query:(language:kuery,query:''),uiState:(),vis:(aggs:!((enabled:!t,id:'1',params:(),schema:metric,type:count),(enabled:!t,id:'2',params:(filters:!((input:(language:kuery,query:'sentiment:POSITIVE'),label:Positive),(input:(language:kuery,query:'sentiment:NEGATIVE'),label:Negative),(input:(language:kuery,query:'sentiment:NEUTRAL'),label:Neutral))),schema:segment,type:filters)),params:(addLegend:!t,addTimeMarker:!f,addTooltip:!t,categoryAxes:!((id:CategoryAxis-1,labels:(filter:!f,rotate:0,show:!t,truncate:200),position:left,scale:(type:linear),show:!t,style:(),title:(),type:category)),dimensions:(x:(accessor:0,aggType:filters,format:(),params:()),y:!((accessor:1,aggType:count,format:(id:number),params:()))),grid:(categoryLines:!f),labels:(),legendPosition:right,seriesParams:!((data:(id:'1',label:Count),drawLinesBetweenPoints:!t,mode:normal,show:!t,showCircles:!t,type:histogram,valueAxis:ValueAxis-1)),times:!(),type:histogram,valueAxes:!((id:ValueAxis-1,labels:(filter:!t,rotate:75,show:!t,truncate:100),name:LeftAxis-1,position:bottom,scale:(mode:normal,type:linear),show:!t,style:(),title:(text:Count),type:value))),title:'',type:horizontal_bar))>

**Important metrics:**

1. Positive : 21,002
2. Negative : 212,798
3. Neutral : 26,930

**Analysis :** The overwhelming number of tweets are negative in sentiment followed by neutral and the last is positive.



The above screenshot shows **Positive Negative Neutral percentages in last 5 hours**

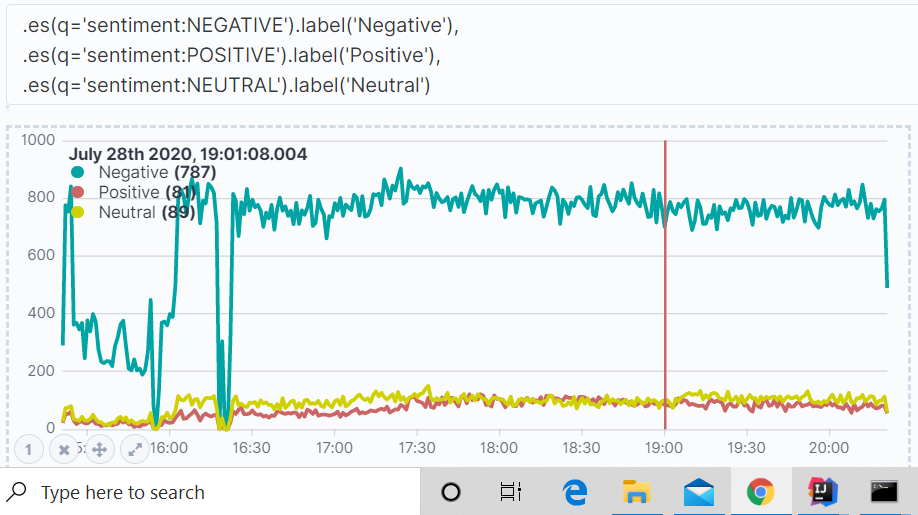
URI to recreate the visualization:

<http://localhost:5601/app/kibana#/visualize/create?type=pie&indexPattern=faa425e0-d089-11ea-9d8c-6f044db2969a&_g=(filters:!(),refreshInterval:(pause:!t,value:0),time:(from:now-15h,to:now))&_a=(filters:!(),linked:!f,query:(language:kuery,query:''),uiState:(),vis:(aggs:!((enabled:!t,id:'1',params:(),schema:metric,type:count),(enabled:!t,id:'2',params:(field:sentiment.keyword,size:3),schema:segment,type:significant_terms)),params:(addLegend:!t,addTooltip:!t,dimensions:(metric:(accessor:0,aggType:count,format:(id:number),params:())),isDonut:!t,labels:(last_level:!t,show:!f,truncate:100,values:!t),legendPosition:right,type:pie),title:'',type:pie))>

**Important metrics:**

1. Positive : 8.16% (21,282)
2. Negative : 81.4% (212,181)
3. Neutral : 10.43% (27,198)

**Analysis :** Most of the tweets are negative in sentiment followed by neutral and the last is positive with only 8%.

****

The above screenshot shows **Minute-by-minute graph of Positive Negative Neutral percentages in last 5 hours**

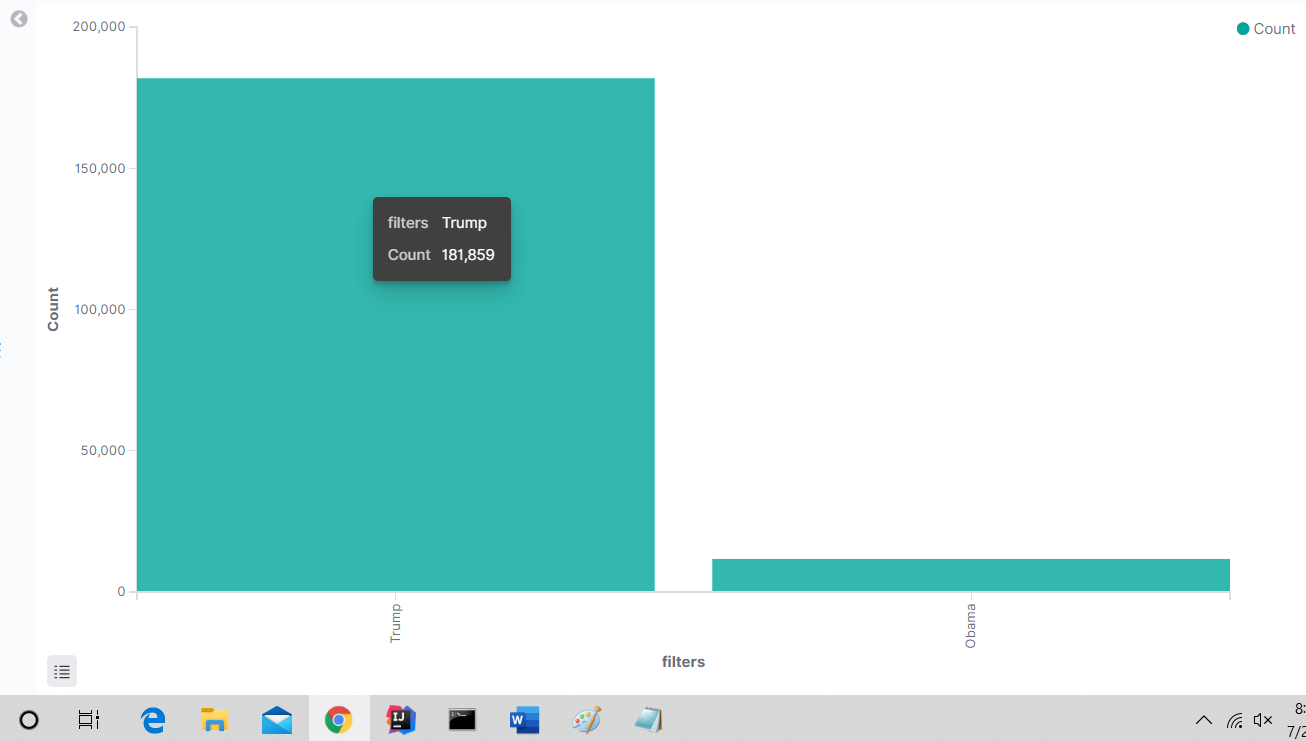
URI to recreate the visualization:

http://localhost:5601/app/timelion#?\_g=(refreshInterval:(pause:!t,value:0),time:(from:now-5h,to:now))&\_a=(columns:2,interval:'1m',rows:2,selected:0,sheet:!('.es(q%3D!'sentiment:NEGATIVE!').label(!'Negative!'),%0A.es(q%3D!'sentiment:POSITIVE!').label(!'Positive!'),%0A.es(q%3D!'sentiment:NEUTRAL!').label(!'Neutral!')'))

**Important metrics:**

1. Max tweets per min : 900
2. Time period : 3:00 PM to 8:00 PM

**Analysis :** This graph also reconfirms our previous analysis at a more granular level that most of the tweets are negative, followed by neutral and then positive.



The above screenshot shows **Absolute count of tweets with mention of Trump as compared to Obama in last 5 hours**

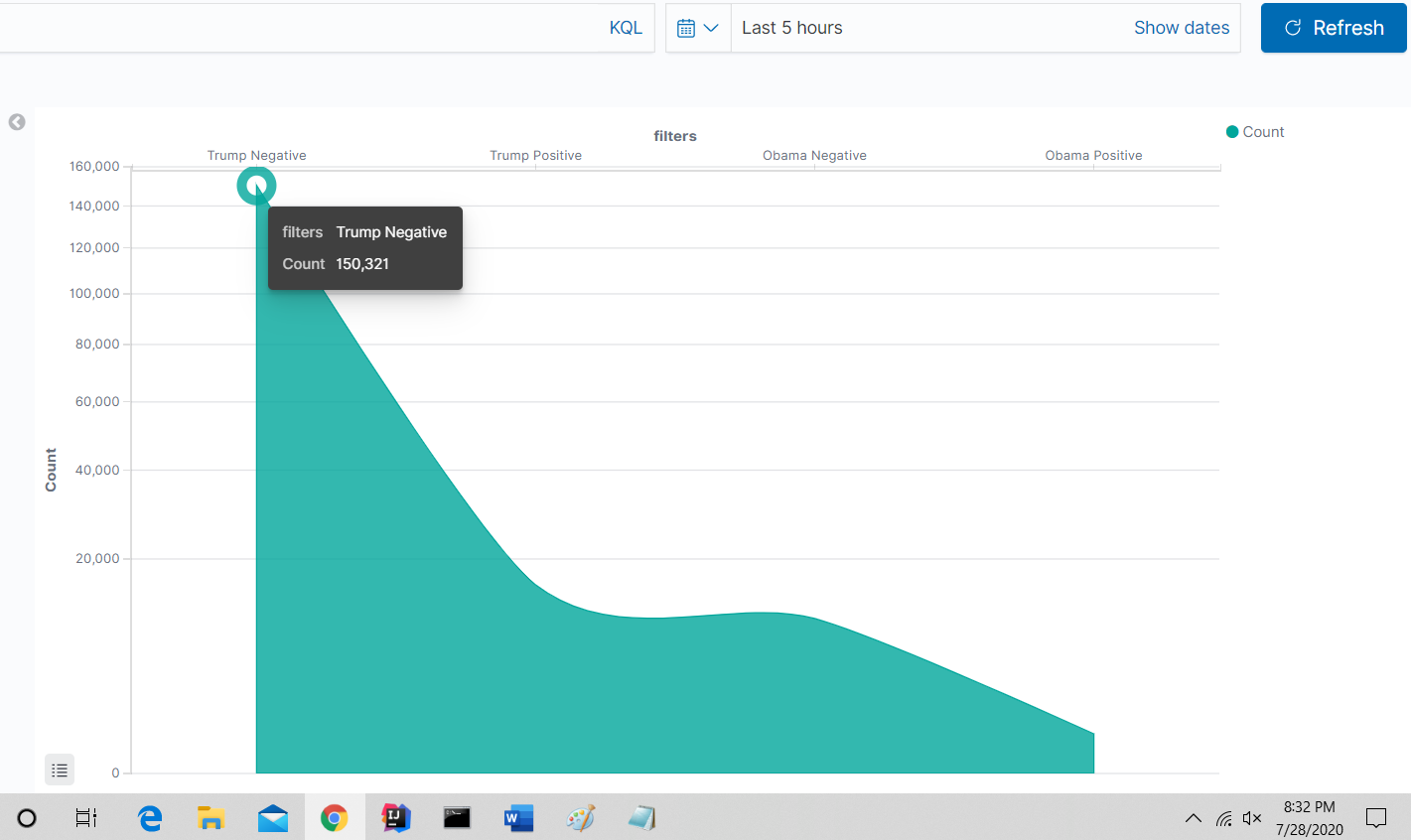
URI to recreate the visualization:

[http://localhost:5601/app/kibana#/visualize/create?type=histogram&indexPattern=faa425e0-d089-11ea-9d8c-6f044db2969a&\_g=(refreshInterval:(pause:!t,value:0),time:(from:now-5h,to:now))&\_a=(filters:!(),linked:!f,query:(language:kuery,query:''),uiState:(),vis:(aggs:!((enabled:!t,id:'1',params:(),schema:metric,type:count),(enabled:!t,id:'2',params:(filters:!((input:(language:kuery,query:'sentence:\*Trump\*'),label:Trump),(input:(language:kuery,query:'sentence:\*Obama\*'),label:Obama))),schema:segment,type:filters)),params:(addLegend:!t,addTimeMarker:!f,addTooltip:!t,categoryAxes:!((id:CategoryAxis-1,labels:(filter:!t,show:!t,truncate:100),position:bottom,scale:(type:linear),show:!t,style:(),title:(),type:category)),dimensions:(x:(accessor:0,aggType:filters,format:(),params:()),y:!((accessor:1,aggType:count,format:(id:number),params:()))),grid:(categoryLines:!f),labels:(show:!f),legendPosition:right,seriesParams:!((data:(id:'1',label:Count),drawLinesBetweenPoints:!t,mode:stacked,show:true,showCircles:!t,type:histogram,valueAxis:ValueAxis-1)),thresholdLine:(color:%2334130C,show:!f,style:full,value:10,width:1),times:!(),type:histogram,valueAxes:!((id:ValueAxis-1,labels:(filter:!f,rotate:0,show:!t,truncate:100),name:LeftAxis-1,position:left,scale:(mode:normal,type:linear),show:!t,style:(),title:(text:Count),type:value))),title:'',type:histogram))](http://localhost:5601/app/kibana#/visualize/create?type=histogram&indexPattern=faa425e0-d089-11ea-9d8c-6f044db2969a&_g=(refreshInterval:(pause:!t,value:0),time:(from:now-5h,to:now))&_a=(filters:!(),linked:!f,query:(language:kuery,query:''),uiState:(),vis:(aggs:!((enabled:!t,id:'1',params:(),schema:metric,type:count),(enabled:!t,id:'2',params:(filters:!((input:(language:kuery,query:'sentence:*Trump*'),label:Trump),(input:(language:kuery,query:'sentence:*Obama*'),label:Obama))),schema:segment,type:filters)),params:(addLegend:!t,addTimeMarker:!f,addTooltip:!t,categoryAxes:!((id:CategoryAxis-1,labels:(filter:!t,show:!t,truncate:100),position:bottom,scale:(type:linear),show:!t,style:(),title:(),type:category)),dimensions:(x:(accessor:0,aggType:filters,format:(),params:()),y:!((accessor:1,aggType:count,format:(id:number),params:()))),grid:(categoryLines:!f),labels:(show:!f),legendPosition:right,seriesParams:!((data:(id:'1',label:Count),drawLinesBetweenPoints:!t,mode:stacked,show:true,showCircles:!t,type:histogram,valueAxis:ValueAxis-1)),thresholdLine:(color:%2334130C,show:!f,style:full,value:10,width:1),times:!(),type:histogram,valueAxes:!((id:ValueAxis-1,labels:(filter:!f,rotate:0,show:!t,truncate:100),name:LeftAxis-1,position:left,scale:(mode:normal,type:linear),show:!t,style:(),title:(text:Count),type:value))),title:'',type:histogram)))

**Important metrics:**

1. Trump : 181,859
2. Obama : 11,703

**Analysis :** Trump is much more famous (or infamous) among twitter users as compared to Obama.



The above screenshot shows **Absolute count of tweets with mention of Trump as compared to Obama with sentiment filter in last 5 hours**

URI to recreate the visualization:

[http://localhost:5601/app/kibana#/visualize/create?type=histogram&indexPattern=faa425e0-d089-11ea-9d8c-6f044db2969a&\_g=(refreshInterval:(pause:!t,value:0),time:(from:now-5h,to:now))&\_a=(filters:!(),linked:!f,query:(language:kuery,query:''),uiState:(),vis:(aggs:!((enabled:!t,id:'1',params:(),schema:metric,type:count),(enabled:!t,id:'2',params:(filters:!((input:(language:kuery,query:'sentiment:POSITIVE%20and%20sentence:\*Trump\*'),label:'Trump%20Positive'),(input:(language:kuery,query:'sentiment:POSITIVE%20and%20sentence:\*Obama\*'),label:'Obama%20Positive'),(input:(language:kuery,query:'sentiment:NEGATIVE%20and%20sentence:\*Trump\*'),label:'Trump%20Negative'),(input:(language:kuery,query:'sentiment:NEGATIVE%20and%20sentence:\*Obama\*'),label:'Obama%20Negative'))),schema:segment,type:filters)),params:(addLegend:!t,addTimeMarker:!f,addTooltip:!t,categoryAxes:!((id:CategoryAxis-1,labels:(filter:!t,rotate:0,show:!t,truncate:100),position:top,scale:(type:linear),show:!t,style:(),title:(),type:category)),dimensions:(x:(accessor:0,aggType:filters,format:(),params:()),y:!((accessor:1,aggType:count,format:(id:number),params:()))),grid:(categoryLines:!f,valueAxis:ValueAxis-1),labels:(show:!f),legendPosition:right,orderBucketsBySum:!t,seriesParams:!((data:(id:'1',label:Count),drawLinesBetweenPoints:!t,interpolate:cardinal,mode:stacked,show:true,showCircles:!t,type:area,valueAxis:ValueAxis-1)),thresholdLine:(color:%2334130C,show:!f,style:full,value:10,width:1),times:!(),type:histogram,valueAxes:!((id:ValueAxis-1,labels:(filter:!f,rotate:0,show:!t,truncate:100),name:LeftAxis-1,position:left,scale:(mode:normal,type:'square%20root'),show:!t,style:(),title:(text:Count),type:value))),title:'',type:histogram))](http://localhost:5601/app/kibana#/visualize/create?type=histogram&indexPattern=faa425e0-d089-11ea-9d8c-6f044db2969a&_g=(refreshInterval:(pause:!t,value:0),time:(from:now-5h,to:now))&_a=(filters:!(),linked:!f,query:(language:kuery,query:''),uiState:(),vis:(aggs:!((enabled:!t,id:'1',params:(),schema:metric,type:count),(enabled:!t,id:'2',params:(filters:!((input:(language:kuery,query:'sentiment:POSITIVE%20and%20sentence:*Trump*'),label:'Trump%20Positive'),(input:(language:kuery,query:'sentiment:POSITIVE%20and%20sentence:*Obama*'),label:'Obama%20Positive'),(input:(language:kuery,query:'sentiment:NEGATIVE%20and%20sentence:*Trump*'),label:'Trump%20Negative'),(input:(language:kuery,query:'sentiment:NEGATIVE%20and%20sentence:*Obama*'),label:'Obama%20Negative'))),schema:segment,type:filters)),params:(addLegend:!t,addTimeMarker:!f,addTooltip:!t,categoryAxes:!((id:CategoryAxis-1,labels:(filter:!t,rotate:0,show:!t,truncate:100),position:top,scale:(type:linear),show:!t,style:(),title:(),type:category)),dimensions:(x:(accessor:0,aggType:filters,format:(),params:()),y:!((accessor:1,aggType:count,format:(id:number),params:()))),grid:(categoryLines:!f,valueAxis:ValueAxis-1),labels:(show:!f),legendPosition:right,orderBucketsBySum:!t,seriesParams:!((data:(id:'1',label:Count),drawLinesBetweenPoints:!t,interpolate:cardinal,mode:stacked,show:true,showCircles:!t,type:area,valueAxis:ValueAxis-1)),thresholdLine:(color:%2334130C,show:!f,style:full,value:10,width:1),times:!(),type:histogram,valueAxes:!((id:ValueAxis-1,labels:(filter:!f,rotate:0,show:!t,truncate:100),name:LeftAxis-1,position:left,scale:(mode:normal,type:'square%20root'),show:!t,style:(),title:(text:Count),type:value))),title:'',type:histogram)))

**Important metrics:**

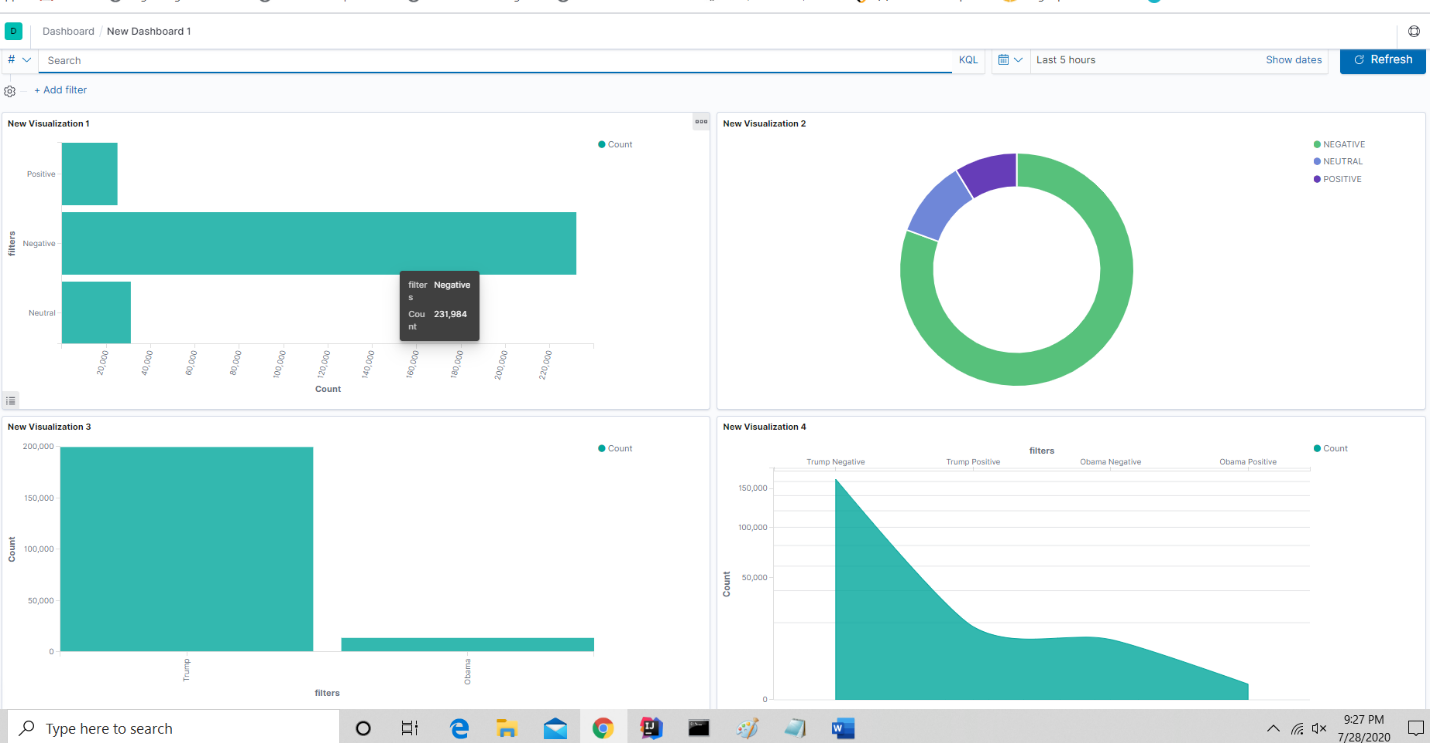
1. Trump Negative : 150,321
2. Trump Positive : 15,475
3. Obama Negative : 10,438
4. Obama Positive : 686

**Analysis :** This graph confirms all our earlier biases.

1. Majority of the tweets on twitter are negative on either topic Trump as well as Obama.
2. Trump is much more in twitter’s limelight as compared to Obama.
3. In absolute terms, Trump has more positive tweets about him that Obama has negative tweets about him.

## Dashboard

We can see all the related Visualizations in Kibana in the form of a dashboard. Please see all the above visualization in the dashboard below:



**Part 2 - Analyzing Social Networks using GraphX**

# Software Requirements

1. java version 1.8.0\_261
2. scala 2.11.8
3. Spark 2.4.0 built for Hadoop 2.7.3
4. graphframes-0.7.0-spark2.4-s\_2.11 (provided as a jar)

# Dataset

1. The Dataset is taken from the url https://snap.stanford.edu/data/web-Google.html. web pages and the hyperlinks are represented by Nodes and directed edges respectively. Google released this data in 2002.

# How to Execute?

1. Open the project in the IntelliJ platform.
2. In the edit Configuration give arguments <inputDirectory> <outputDirectory> as input\web-Google.txt output
3. Result of 5 questions will be in 5 separate files.

# Summary

1. Find the top 5 nodes with the highest outdegree and find the count of the number of outgoing edges in each

(id,outgoing edges count)

(49,622652)

(50,621138)

(53,620638)

(52,620519)

(54,619552)

1. Find the top 5 nodes with the highest indegree and find the count of the number of incoming edges in each

(id,incoming edges count)

(48,557240)

(49,533668)

(53,505196)

(56,504335)

(51,502987)

1. Calculate PageRank for each of the nodes and output the top 5 nodes with the highest PageRank values. You are free to define the threshold parameter.

(id,pagerank)

(9,1.4216867438886687)

(48,1.3349187391231192)

(49,1.2185617591904208)

(53,1.0302195604227466)

(56,1.0289510335341745)

1. Run the connected components algorithm on it and find the top 5 components with the largest number of nodes.

(component,count)

(35,32)

(32,32)

(54,9)

(48,9)

(57,9)

1. Run the triangle counts algorithm on each of the vertices and output the top 5 vertices with the largest triangle count. In case of ties, you can randomly select the top 5 vertices.

(vertex id, number of triangles)

(54,44)

(48,44)

(57,44)

(52,44)

(55,44)