ReadMe- Assignment Part 1

Running the jar file and displaying sentiments on Kafka consumer topic

Step 1:

Download Kafka using the following command:
tar -xzf kafka_2.12-2.2.0.tgz

After installing Kafka,
Go to kafka_2.12-2.2.0 bin folder and run this command

sh zookeeper-server-start.sh "path-to-kafka-server"/kafka_2.12-2.2.0/config/zookeeper.properties

This will start the Zookeeper server

Step 2:

Stay in the bin folder and start the kafka server using the following

sh kafka-server-start.sh /Users/yamadane/Desktop/Spring_2019/Big_Data/kafka_2.12-2.2.0/config/server.properties

This will start the Kafka server

Step 3:

Create a topic

Stay in the bin folder and create a Kafka topic using the following command:
I have used twitter_sent as the topic in my assignment

sh kafka-topics.sh --create --bootstrap-server localhost:9092 --replication-factor 1 --partitions 1 --topic twitter_sent

Step 4:

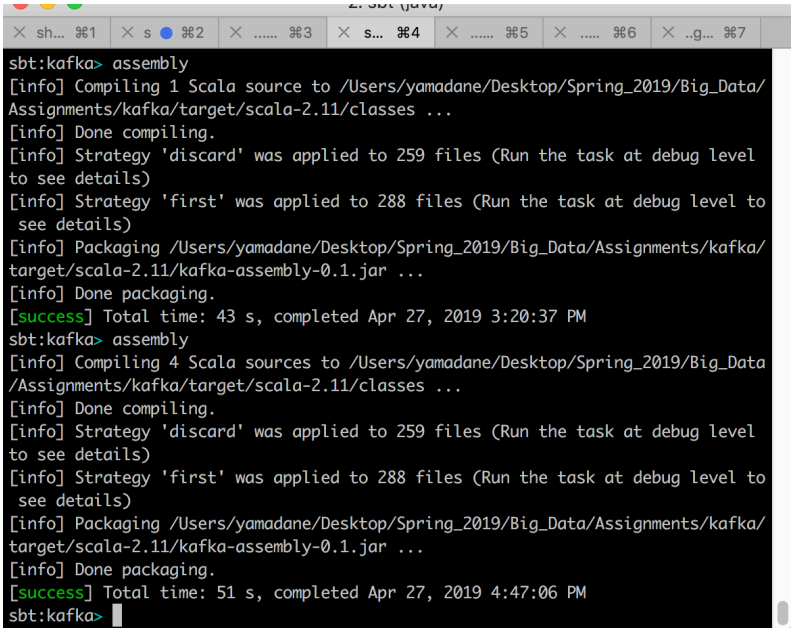
Unzip the project kapka.zip

Note: I have deleted the fat jar as it was of around 250 MB size

Go to the the home directory of the kafka project in your terminal and do sbt

Make sure you have sbt installed on your PC/MAC. If not you can do a brew install sbt

After the sbt console opens , type the command assembly. It will generate a fat jar which consists of all the scala classes and dependencies



Step 5:

Running the jar:

Step 1:
  Install spark on your local machine. On Mac, you could do this using brew install apache-spark

Command to run the jar

spark-submit --packages org.apache.spark:spark-streaming_2.11:1.6.2 --class TwitterProducer “location-of-the-fatjar-file” kafka-topic filter < Twitter Consumer Key> < Twitter Consumer Secret> < Twitter Access Token> < Twitter Access Token Secret>

Arguments:

1) Kafka topic. For this project it is twitter_sent
2) Twitter filter- The hashtag which you want your sentimental analysis to be done. For me it is trump
3) Twitter Consumer Key (API key)
4) Twitter Consumer Secret(API Secret Key)

5) Twitter Access Token(Access Token)
6) Twitter Access Token Secret(AccessTokenSecret)

Note : It is very importance to run **org.apache.spark:spark-streaming_2.11:1.6.2** version as logging is supported only in that version. It wont work for other sparkstreaming versions. Atleast not that I know of.

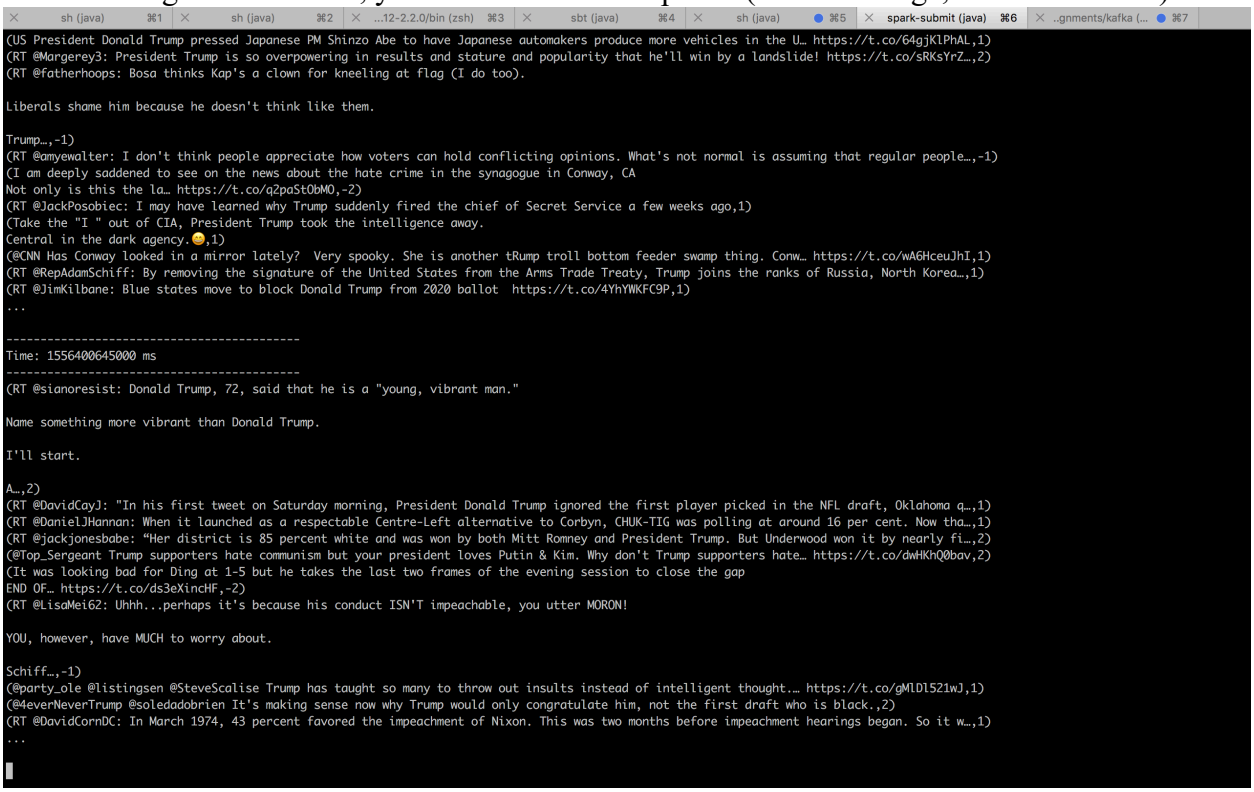
The command which I ran:

spark-submit --packages org.apache.spark:spark-streaming_2.11:1.6.2 --class TwitterProducer /Users/yamadane/Desktop/Spring_2019/Big_Data/Assignments/kafka/target/scala-2.11/kafka-assembly-0.1.jar twitter_sent "trump" Twitter Consumer Key Twitter Consumer Secret Twitter Access Token Twitter Access Token Secret

I am hiding the twitter keys for security purposes

This will be the output on the terminal after running this command. The jar will fetch tweets related to trump in a window of 15 seconds , perform sentimental analysis on the tweets and sends the sentiment score to the kafka consumer.

After running the command, you should see the output as (tweet message, sentiment value) .



Simultaneously, the sentiment score is being sent to the topic twitter_sent

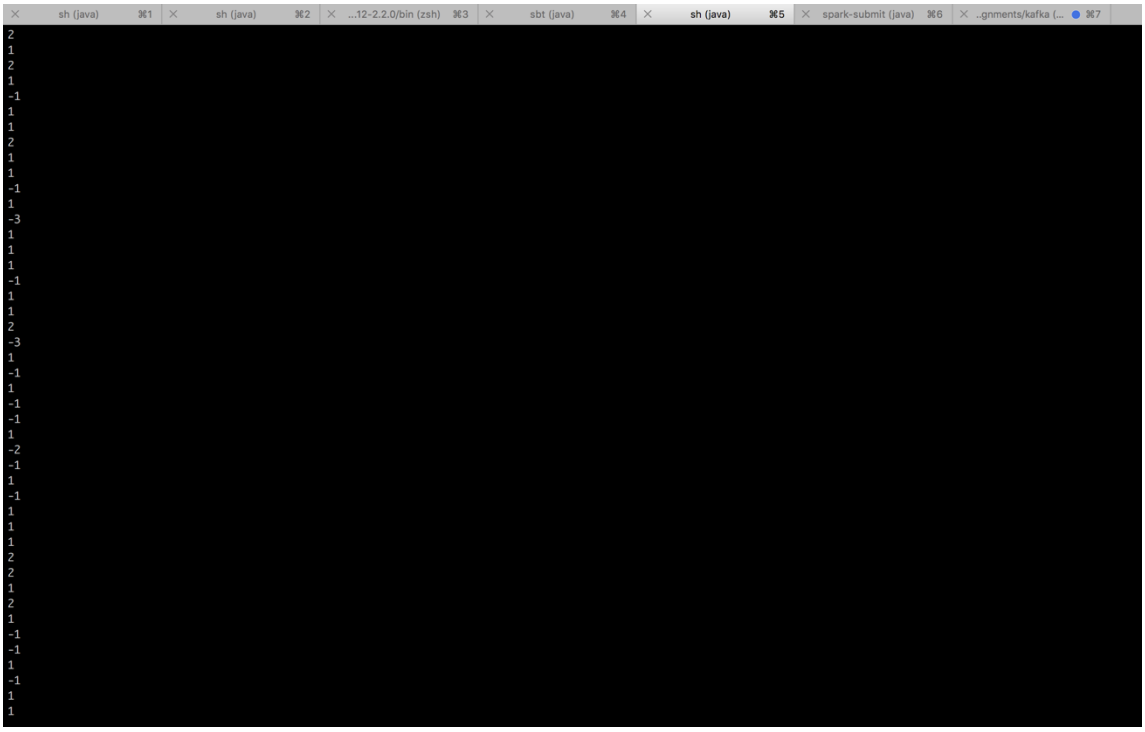
To see the data on the kafka topic run this command:

1) Go to kafka_2.12-2.2.0 bin folder and run this command
   sh kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic “topic_name” --from-beginning

   Note: It is important to run zookeeper and kafka server during the entire process

   sh kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic twitter_sent --from-beginning

   This should be the output after running the commands. These are nothing but the twitter sentiment values



   You can sent the twitter messages and sentiment together if you want to by changing line 92 from
   val data = record._2.toString to record.toString

   Mostly the sentiments have a range spread from -2 to 2, where -2 is really negative, -1 is negative, 1 is positive, and 2 is really positive.

   Sentiment scores are generally a better measure than just positive or negative sentiments.

*** Important**
If any of the steps give an error ,
I have uploaded the jar on box. This is the link. This jar has my twitter credentials embedded in the code so you won't need to give the twitter arguments
https://utdallas.box.com/s/3fhndjcpe2upgk00ckg7kqfu1vgl3bz0

You can run this jar using
spark-submit --packages org.apache.spark:spark-streaming_2.11:1.6.2 --class TwitterProducer /jar-location/kafka-assembly-0.1.jar twitter_sent "trump"

This is the command which I did
spark-submit --packages org.apache.spark:spark-streaming_2.11:1.6.2 --class TwitterProducer /Users/yamadane/Desktop/Spring_2019/Big_Data/Assignments/kafka/target/scala-2.11/kafka-assembly-0.1.jar twitter_sent "trump"

Visualising the data

Step 1:

Download elastic search , kibana and logstash using the link provided in the assignment details

Step 2:
Go to elastic search bin directory and run this command
sh elasticsearch

Step 3:
Go to kibana directory and run this command:
sh kibana

Step 4:
Create a file named logstash-simple.conf in the bin folder of logstash and paste the following:

```
input {
kafka {
bootstrap_servers => "localhost:9092"
topics => ["twitter_sent"]
}
}
```

```
output {
elasticsearch {
hosts => ["localhost:9200"]
index => "twitter_sent-index"
}
}
```

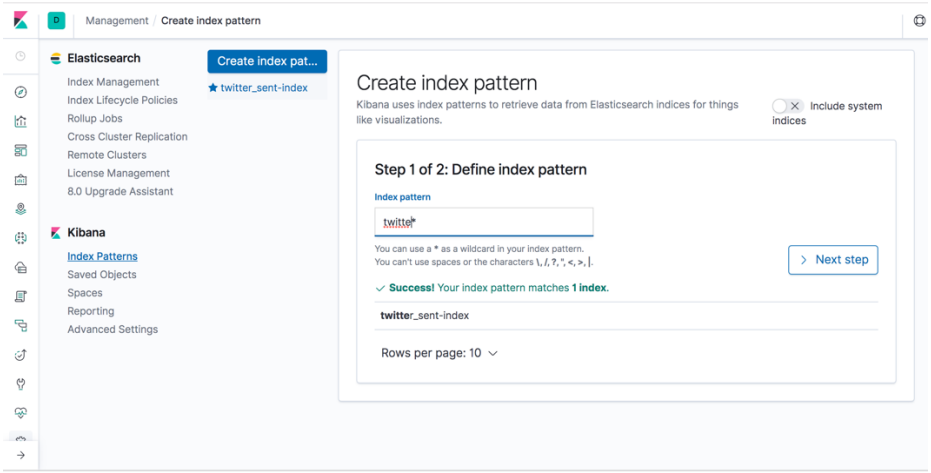
Then run this command in the bin directory:

sh logstash -f logstash-simple.conf

If everything goes fine you should find your topic on Kibana:

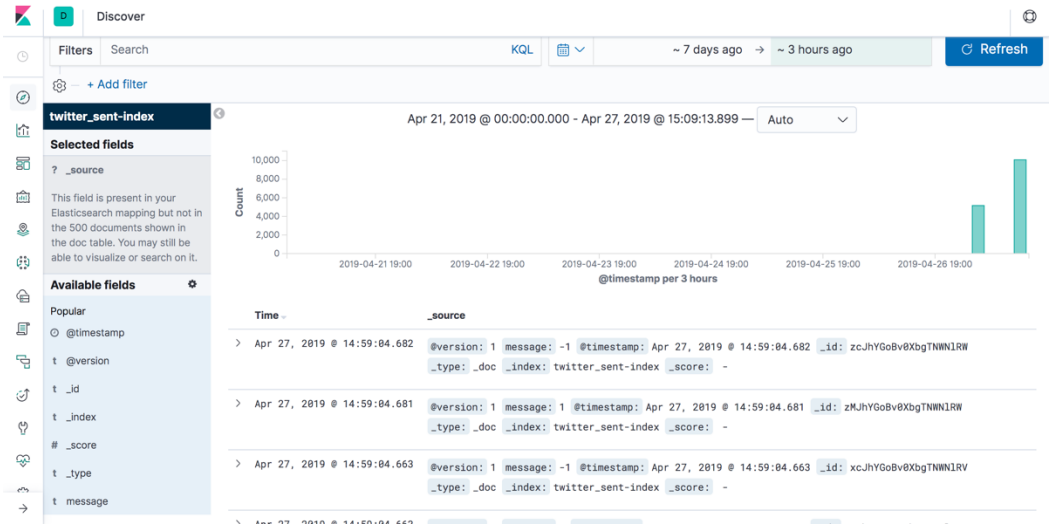
Kibana

Go to Kibana dashboard, find connect to your elasticsearch engine and select management. If everything is set up correctly you would be able to find your index pattern:



Go to the next step and you can create your visualisations:

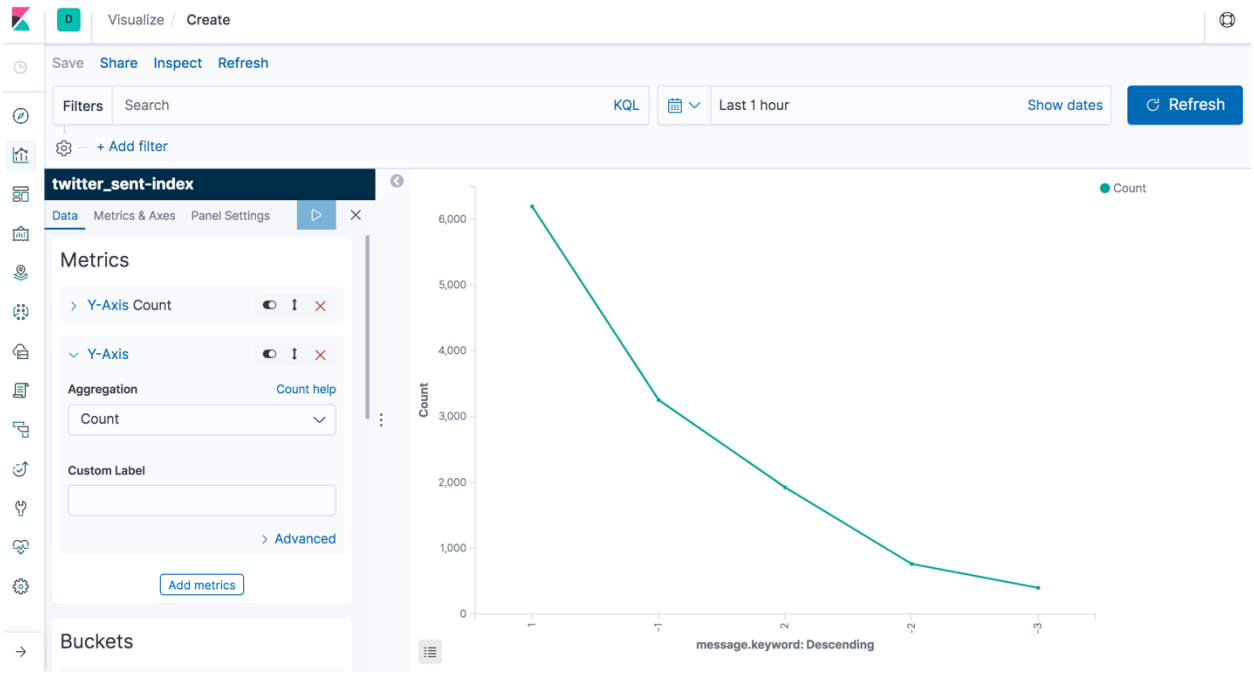
You can even see your data:



Visualisation 1:

This visualization shows the sentiment value of the tweets containing the keyword trump after 30 minutes:

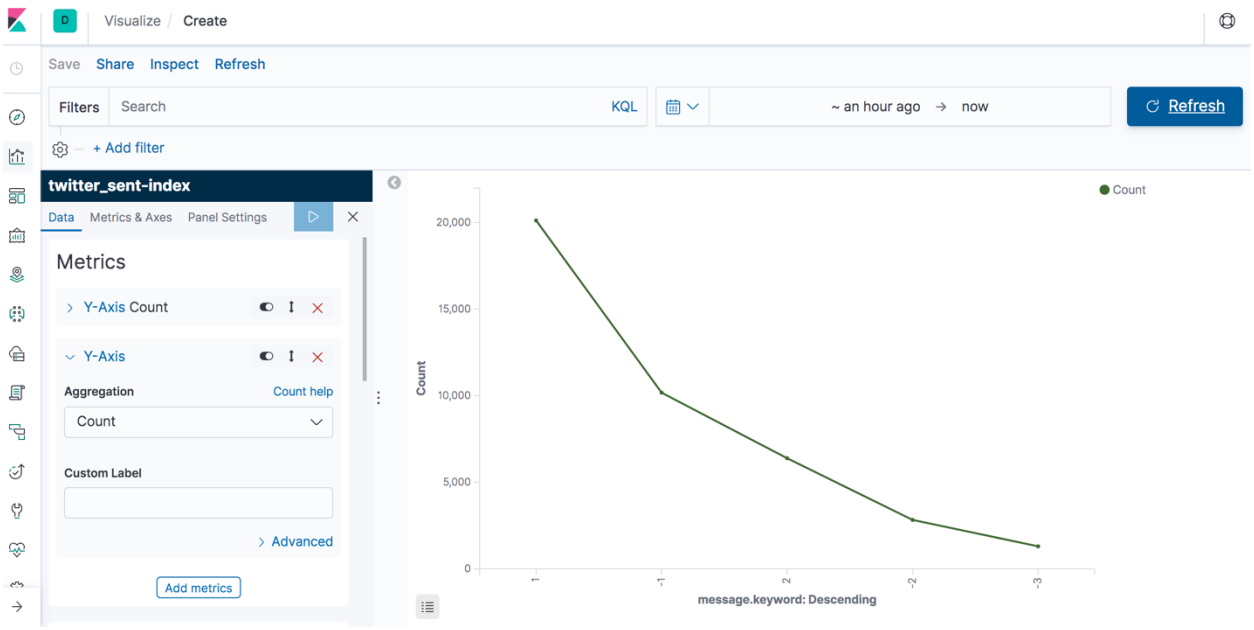
The x-axis contains the message i..e the sentiment value 1, -1 ,2 , -2, 3. The y-axis shows the sentiment count sorted in the graph



Conclusion:

We can clearly see that the sentiment count is positive with around 6000 positive tweets about trump in the past 30 minutes.

Visualisation 2:
The same visulation is now done for around an hour and half to see if there are any changes in the trend



Conclusion:
The trend is still positive towards trump after an hour and half. This sentimental analysis will be useful during elections to see the attitude of people towards a particular electoral candidate.

I have also attached the csv files of the count generated by Kibana.

Sources:
https://github.com/vaquarkhan/spark-twitter-stream-sentiment-analysis

https://kafka.apache.org/quickstart