TWITTER SENTIMENT ANALYSIS USING SPARK STREAMING

BY
Manthiramoorthy Cheranthian
Akshaya Sivakumar Karunambika

Project overview

Sentiment Analysis - contextual mining of text which identifies and extracts subjective information to understand the social sentiment for a topic.

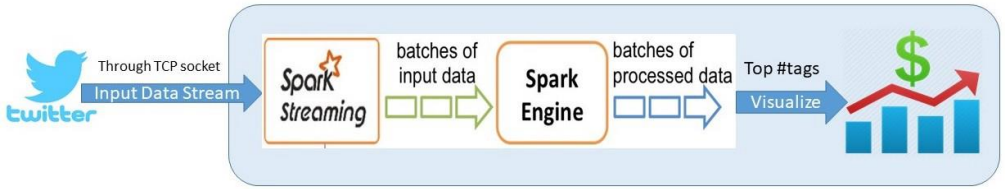
With people looking to social media to voice their immediate opinions and feedback, we decided to work on **Twitter** data.

In order to get real-time live sentiment for business reviews, we used **Spark** streaming.

We focused on comparing the sentiment scores of live streaming tweets for two commerce giants **Costco and Walmart** in the United States.

Spark Streaming using TCP Socket

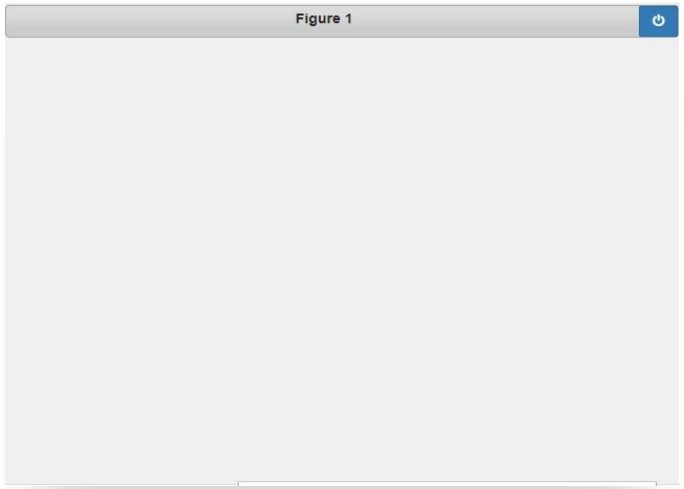
- Spark Streaming is an extension of the core Spark API to process real-time data from source like TCP Socket , Kafka, Flume etc.
- Create credentials for accessing the Twitter API
- App host socket connection is set up for Spark to transfer the extracted tweets – Tweets are extracted only for the United States
- Streaming client is created to receive tweets related to Walmart and Costco in intervals of 3 seconds
- Text analysis is performed on received tweets after pre-processing steps and sentiment score is reflected real-time in a graph

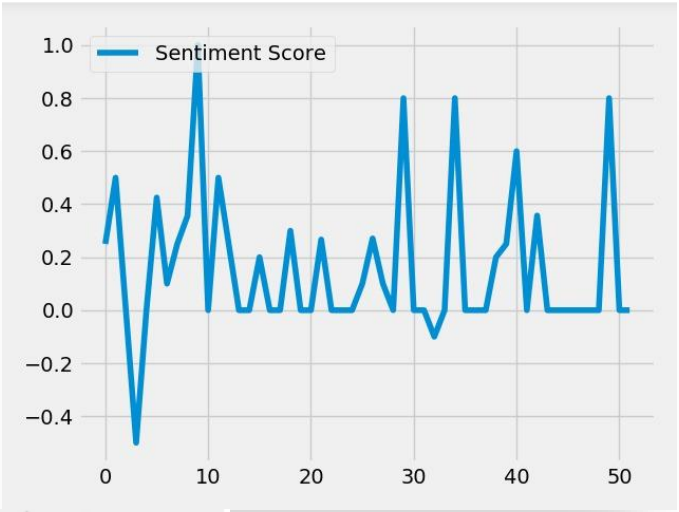


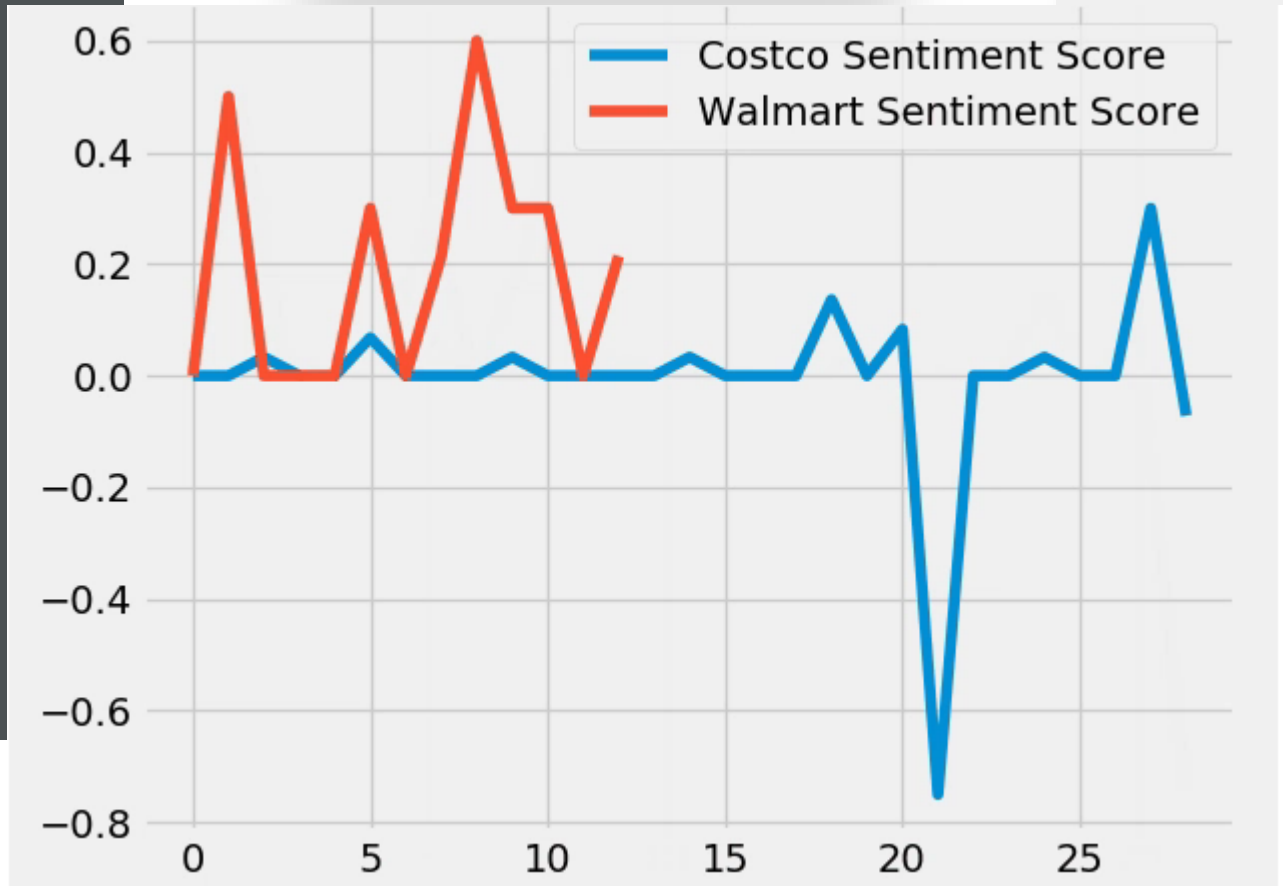

```
Connected... Starting getting tweets.
https://stream.twitter.com/1.1/statuses/filter.json?language=en&track=Costco <Response [200]>
Tweet Text: RT @SEDLAW15: Quote of the year: “If you expect elementary school children to endure the traum
r drills for your freedoms,…
------------------------------------------
Tweet Text: RT @chipfranklin: Doctors/nurses can’t come home &amp;amp; hug their kids. They stay at work &amp;
patients, their colleagues, their…
------------------------------------------
Tweet Text: RT @leahjdouglas: NEW: A worker at the Nebraska plant that produces chicken for Costco has die
workers are sick. The plant…
------------------------------------------
Tweet Text: Thank Gawd, too many people in here anyway, head to dollar tree buh-bye
------------------------------------------
Tweet Text: RT @nursekelsey: “If you expect elementary school children to endure the trauma of active shoo
r freedoms, you can wear a…
------------------------------------------
Tweet Text: RT @nmeyersohn: A worker at Costco's poultry plant in Nebraska has died from the coronavirus.
lant have tested positive.…
------------------------------------------
Tweet Text: @UP10516706 @chucklorrefans Btw, you need to know your constitutional rights.  They protect yo
ent… https://t.co/XI426mmpJe
------------------------------------------
Tweet Text: Costco, Kroger warn of limited supply https://t.co/IMauotUkWF https://t.co/tAm6r8kojP
```

```
Waiting for TCP connection...
Connected... Starting getting tweets.
https://stream.twitter.com/1.1/statuses/filter.json?language=en&track=Walmart
Tweet Text: making a walmart list for pickup and decided fuck it i'ma order a
------------------------------------------
Tweet Text: @GregRaths @OCGOPNews #Virus comes from #China, world goes on #loc
art… https://t.co/auSYYO1CaY
------------------------------------------
Tweet Text: RT @JoeySalads: What if Lockdowns caused the virus to spread faste

Democrats don’t allow people to relax on an open-air beach but can al…
------------------------------------------
Tweet Text: RT @Future_SOTH: @BuzzPatterson #Virus comes from #China, world go
ose, #WalMart large coopera…
------------------------------------------
Tweet Text: RT @AmandaPresto: Had to stop at the local Walmart today. It was
s lockdown is. I can go to…
```







Real-time Sentiment Score per tweet

- For each tweet received , Sentiment score is calculated
- The score is then written into a csv file
- The csv file is given as input to the graph to plot the sentiment score as it is received in real-time
- MatPlotLib was used for real-time chart creation

Analysis on Tweet text for Costco and Walmart

For each tweet received , The text is tokenized

All the stop words , special characters are removed

Word vector is created

HashingTF-IDF+ Logistic regression model trained on Sentiment140 dataset

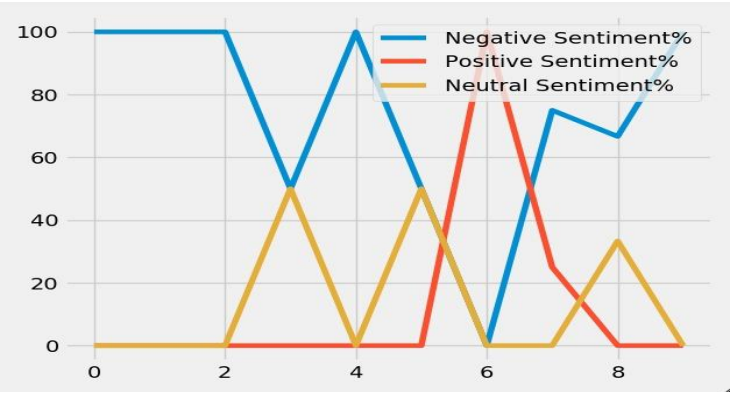
Model is used to classify the tweets into 0-negative,2-neutral,3-positive

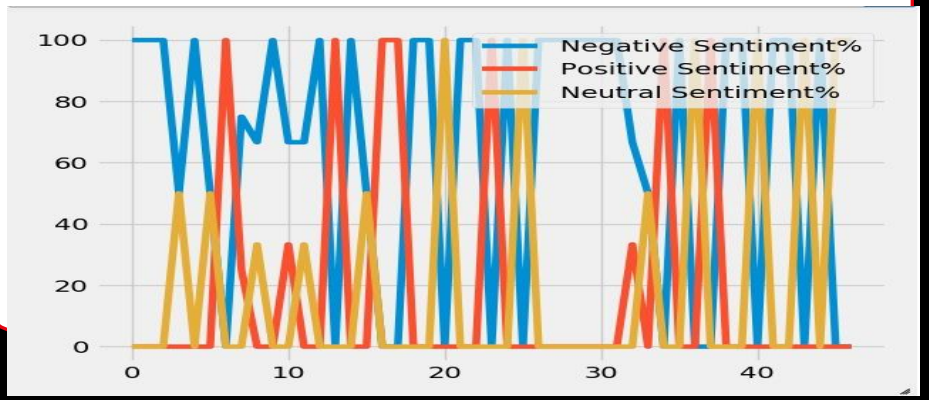
The percentage of negative, neutral, positive tweet received are aggregated

The score is then written into a csv file

The csv file is given as input to the graph to plot the sentiment score as it is received in real-time

```
Negative tweets percentage:% 40.0
Positive tweets percentage:% 20.0
Neutral tweets percentage:% 40.0
+--------------------+----------+
|               tweet|prediction|
+--------------------+----------+
| RT @COsweda: Thread|       2.0|
|      What a shame. |       0.0|
|@Timcast went bac...|       2.0|
|It looks like he'...|       0.0|
|RT @SEDLAW15: Quo...|       4.0|
+--------------------+----------+
```



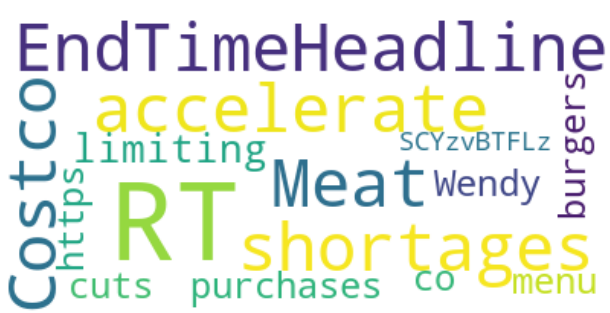


Comparing Sentiment by Region

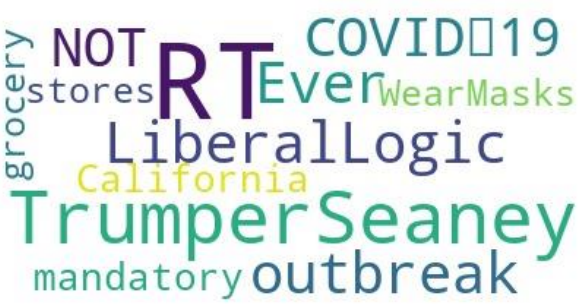
- Filtered tweets for San Francisco and New York
- Generated word cloud for Costco and Walmart for both regions
- Interesting insights in sync with current news

-122.75,36.8,-121.75,37.8	San Francisco
-74,40,-73,41	New York City

Costco New York



Costco San Francisco



Costco to limit meat purchases

NEW YORK CITY (SBG) - For some people, quarantine has come to represent a time of heightened creativity, with plenty of inspiration to be ...

1 day ago

Coronavirus Update: San Francisco Bay Area Costco Shoppers Now Required To Wear Masks; Limits To Meat…

Coronavirus Update: San Francisco Bay Area Costco Shoppers Now Required To Wear Masks; Limits To Meat Purchases. May 4, 2020 at 7:52 ...

1 day ago

Walmart New York

Walmart San Francisco

Insights and enhancements

Live sentiment scores are indicators of customer's reaction for the products and services provided by the enterprises.

Tracking this score will help identify the areas that need to be targeted to achieve higher customer satisfaction.

Focused marketing and promotions will help in increasing the sales

Area wise tracking would help in identifying the regions where the business is performing well and where it needs to be improved.

THANK YOU