stock_comment
About Dataset

Originally we were planning to scrap conversation data from Yahoo finance (https://finance.yahoo.com/quote/AA/community). The conversation data is recent and diverse. But web-scrapping consumes a lot of time. After 20 hours running on local machine, only about 30MB of data has been generated. I also tried to run the script on a CIMS server, but Google chrome runs very slowly inside the virtual box. Thus, considering time and data size, I obtained the Stockwits dataset(about 194.3MB) collected by an Udacity Team as an alternative option. The dataset contains messages from Stockwits(a social media app), and those messages are similar to posts on twitter. This dataset is available in the pulic domain and contains sufficient data. More detailed description can be found here (https://vkontech.com/sentiment-analysis-of-stocktwits-messages-using-lstm-in-pytorch/).

Exploratory data analysis & cleansing

Here, I created a schema for the dataframe, called z.show() to present some rows of the dataset. In total, there are 4 columns, 1548010 rows. Column names and types are shown in the printSchema output.

index	message_body	sentiment
0	$FITB great buy at 26.00...ill wait	2
1	@StockTwits $MSFT	1
2	#STAAnalystAlert for $TDG : Jefferies Maintains with a rating of Hold setting target price at USD 350.00. Our own verdict is Buy http://www.stocktargetadvisor.com/toprating	2
3	$AMD I heard there’s a guy who knows someone who thinks somebody knows something - on StockTwits.	1
4	$AMD reveal yourself!	0

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

```
res3: Int = 4
```

```
root
 |-- index: integer (nullable = true)
 |-- message_body: string (nullable = true)
```

```
 |-- sentiment: integer (nullable = true)
 |-- timestamp: timestamp (nullable = true)
```

stock_comment

```
res5: Long = 1548010
```

Trim strings and add more features

For the message_column: I trimmed it, replaced tab with space, removed “http” link, then extracted $symbol, @usrnameout, #hashtag out and added three more features into the dataframe.

```
trimedDF: org.apache.spark.sql.DataFrame = [index: int, message_body: string ... 2 more fields]
```

```
+-----+--------------------+---------+-------------------+
|index|        message_body|sentiment|          timestamp|
+-----+--------------------+---------+-------------------+
|    0|$FITB great buy a...|        2|2018-07-01 00:00:09|
|    1|   @StockTwits $MSFT|        1|2018-07-01 00:00:42|
|    2|#STAAnalystAlert ...|        2|2018-07-01 00:01:24|
|    3|$AMD I heard ther...|        1|2018-07-01 00:01:47|
|    4|$AMD reveal yours...|        0|2018-07-01 00:02:13|
|    5|$AAPL Why the dro...|        1|2018-07-01 00:03:10|
|    6|$BA bears have 1 ...|       -2|2018-07-01 00:04:09|
|    7|$BAC ok good we&#...|        1|2018-07-01 00:04:17|
|    8|$AMAT - Daily Cha...|        2|2018-07-01 00:08:01|
|    9|$GME 3% drop per ...|       -2|2018-07-01 00:09:03|
|   10|   $SBUX STRONG BUY!|        2|2018-07-01 00:09:26|
|   11|$SNPS short ratio...|       -2|2018-07-01 00:09:36|
|   12|$NFLX price squee...|        2|2018-07-01 00:12:58|
|   13|@DEEPAKM2013 @Nyt...|        2|2018-07-01 00:13:57|
|   14|BREAKOUT Strates...|        2|2018-07-01 00:14:19|
```

index	message_body	sentiment
0	$FITB great buy at 26.00...ill wait	2
1	@StockTwits $MSFT	1
2	#STAAnalystAlert for $TDG : Jefferies Maintains with a rating of Hold setting target price at USD 350.00. Our own verdict is Buy	2
3	$AMD I heard there’s a guy who knows someone who thinks somebody knows something - on StockTwits.	1
4	$AMD reveal yourself!	0
5	$AAPL Why the drop? I warren Buffet taking out	1

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

index	message_body	sentiment	timestamp

stock_comment

0	$FITB great buy at 26.00...ill wait	2	2018-07-01 00:00:09.0
1	@StockTwits $MSFT	1	2018-07-01 00:00:42.0
2	#STAAnalystAlert for $TDG : Jefferies Maintains with a rating of Hold setting target price at USD 350.00. Our own verdict is Buy	2	2018-07-01 00:01:24.0
3	$AMD I heard there's a guy who knows someone who thinks somebody knows	1	2018-07-01 00:01:47.0

Output is truncated to 102400 bytes. Learn more about **ZEPPELIN_INTERPRETER_OUTPUT_LIMIT**

Summary statistics

In this section, I used the z.show(df.describe()) method to present some statistics about each column in the datasent. In the table below, mean of sentiment can be helpful for later processing. Here, mean of sentiment is 0.21 where max is 2 and min is -2. Sentiment values range in [-2, -1, 0, 1, 2]. As the value approch to 0, the sentence is more neutral, positive trend stands for positive sense and vice versa.

I also presented the length of messages in col("message_body") and showed the max and min length value. Min is 0 and max is 356. Since 0 does not make much sence, I removed them as shown below.

summary	index	message_body	sentiment
count	1548010	1548010	1548010
mean	774004.5	null	0.2126065077098985
stddev	446872.1394416605	null	1.0922968019782755
min	0		-2
max	1548009	🏂 Great Friday in the stock market $GERN $AMD $GSAT	2

summary	length
count	1548010
mean	71.941324
stddev	39.253992

min	0
max	396

stock_comment

```
messageDF: Unit = ()
```

```
minValue: Long = 4
```

```
+------+--------------------+---------+-------------------+--------+-----+-------+
| index|        message_body|sentiment|          timestamp|username|stock|hashtag|
+------+--------------------+---------+-------------------+--------+-----+-------+
|587093|$MNK ----okay her...|        1|2018-08-17 18:17:52|        | $MNK|       |
+------+--------------------+---------+-------------------+--------+-----+-------+
```

```
maxValue: Unit = ()
```

remove extreme values

Since there are 4 rows where the message_body is empty, I removed them and presented the decribe() data after removal.

summary	length
count	1548006
mean	71.941509
stddev	39.253873
min	2
max	396

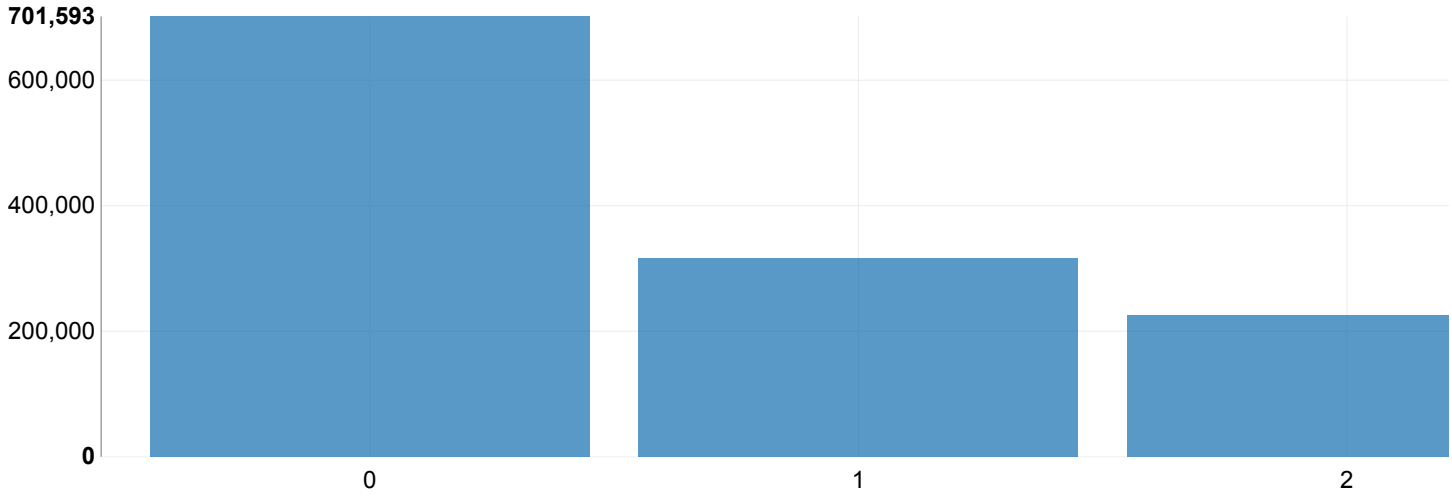
```
cleanedDF: org.apache.spark.sql.Dataset[org.apache.spark.sql.Row] = [index: int, message_body: string
... 5 more fields]
```

perform aggregation

I grouped sentiment column by its score and presented the result in bar chart. From the chart below we can see the value ranges in [-2, -1, 0, 1, 2] where 0 represents neutral, -2 means very negative, -1 means slightly negative, 1 means slightly positive and 2 means very positive. Those sentiments were

hand-labeled by the Udacity team who collected the dataset at the first place. From the result, we can see almost half of the the sentences have 0 as their sentiment score.

stock_comment



output

```
outputPath: String = project/cleanedComments.csv
```