Sentiment Analysis of Twitter Data
for
COVID-19

Goal:

- Study the change in emotion in regard to the pandemic over time
- Find different leaders and celebrities who emerged popular with regard the pandemic
- Find trending events during the pandemic
- Studying an industry standard tool and its efficacy

Tools used for Twitter Data Sentiment Analysis:

- Tweepy- (Data streamed at a gap of 30 minutes for a day)
- SparkML
- SparkSQL
- Sentiment Analysis lib

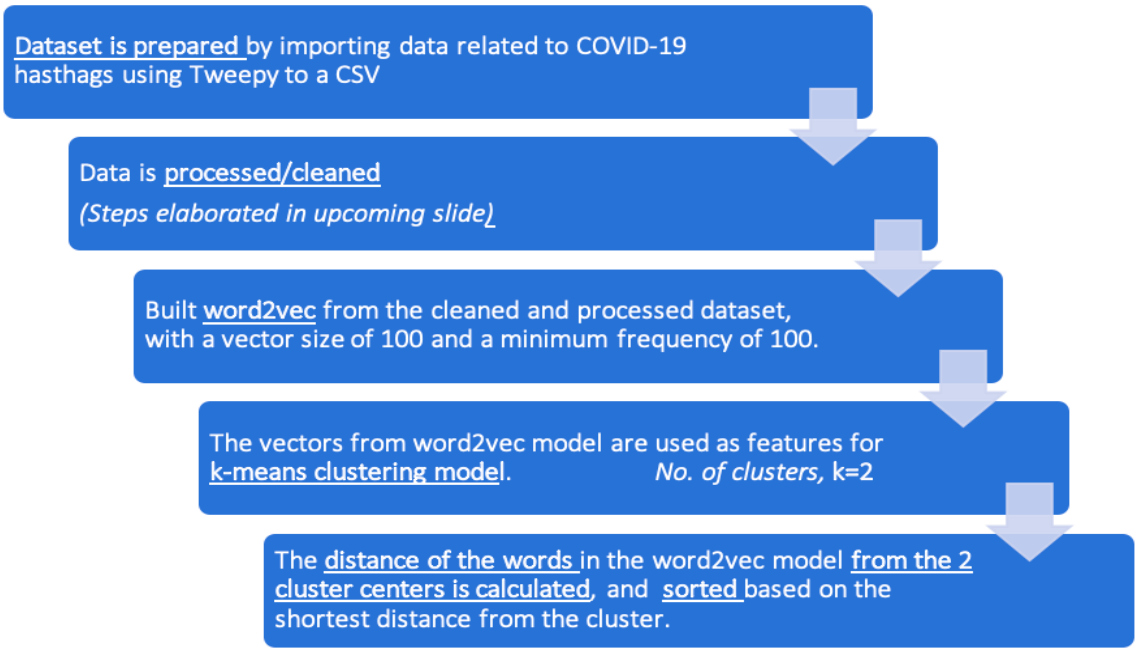
3 Methods Used:

- **SA1**: Unsupervised Clustering
- **SA2**: vaderSentiment Library
- **SA3:** Frequency Analysis + Visualization with Word Cloud

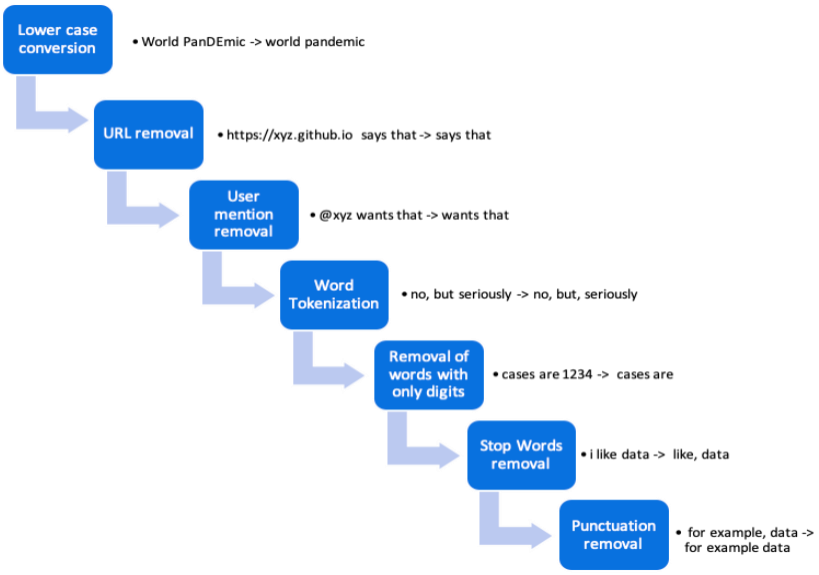
SA 1 : Unsupervised Clustering

- Used K-means Algorithm for Clustering
- 2 clusters formed, one negative and one positive
- Compared change in size of positive and negative clusters over time

SA 1: Unsupervised Clustering Process Flow



SA1: Data Processing Steps for Clustering



SA1: Clustering Results

```
In [37]: word2Vec_df[["word","Distance from Cluster Center 1"]
```

Out[37]:

	word	Distance from Cluster Center 1
214	#bbcqt	0.435903
1895	hello	0.595552
962	#arteta	0.658104
1612	cop	0.676556
592	strike	0.683018
862	ghana	0.702851
159	irony	0.714536
718	awesome	0.717836
928	ring	0.719447
1839	celebrate	0.720573

```
In [38]: word2Vec_df[["word","Distance from Cluster Center 0"]]
```

Out[38]:

	word	Distance from Cluster Center 0
962	#arteta	0.563547
214	#bbcqt	0.583213
862	ghana	0.603011
185	#broadway	0.638844
592	strike	0.658496
2018	king	0.693372
1969	#arsenal	0.701917
362	con	0.713639
2112	lift	0.741400
1219	#ohio	0.757882

```
predictions.select('prediction').groupby('prediction').count().show()
```

```
+----------+------+
|prediction| count|
+----------+------+
|         1|133658|
|         0| 33791|
+----------+------+
```

SA 2 : VaderSentiment Analysis

- VADER (Valence Aware Dictionary and Sentiment Reasoner) tested to see efficiency
- Industry standard tool by MIT
- Specifically attuned to social media
- Intense data cleaning and training not required

SA 2 : VaderSentiment Analysis

In [25]:
```
#top 10 positive sentences
english_tweets_processed.sort(desc("positive_score")).limit(10).show()
```

```
+------------------+--------------------+----------------+--------------+--------------+--------------+-----+
|         status_id|                text|   sentimentscore|negative_score|neutral_score|positive_score|label|
+------------------+--------------------+----------------+--------------+--------------+--------------+-----+
|          So yeah|             yeah?"|[0.0, 0.0, 100.0]|           0.0|           0.0|         100.0|    1|
|1237908954103021568|Confidence inspir...|[0.0, 14.1, 85.9]|           0.0|          14.1|          85.9|    1|
|1238213178393157632|"Please ""GOD"" s...|[0.0, 19.4, 80.6]|           0.0|          19.4|          80.6|    1|
|1238164303787819013|live laugh love #...|[0.0, 20.4, 79.6]|           0.0|          20.4|          79.6|    1|
|1237916361151533058|Wow. Wow. #COVID1...|[0.0, 20.8, 79.2]|           0.0|          20.8|          79.2|    1|
|1238244216456781826|Wow, just wow #Co...|[0.0, 20.8, 79.2]|           0.0|          20.8|          79.2|    1|
|1237906607230550016|Oh dear god #COVI...|[0.0, 21.7, 78.3]|           0.0|          21.7|          78.3|    1|
|1238125020255113216|Stay safe love. #...|[0.0, 22.0, 78.0]|           0.0|          22.0|          78.0|    1|
|1238240121993953281|share share share...|[0.0, 23.3, 76.7]|           0.0|          23.3|          76.7|    1|
|1237910599222394880|Stay safe friends...|[0.0, 25.0, 75.0]|           0.0|          25.0|          75.0|    1|
+------------------+--------------------+----------------+--------------+--------------+--------------+-----+
```

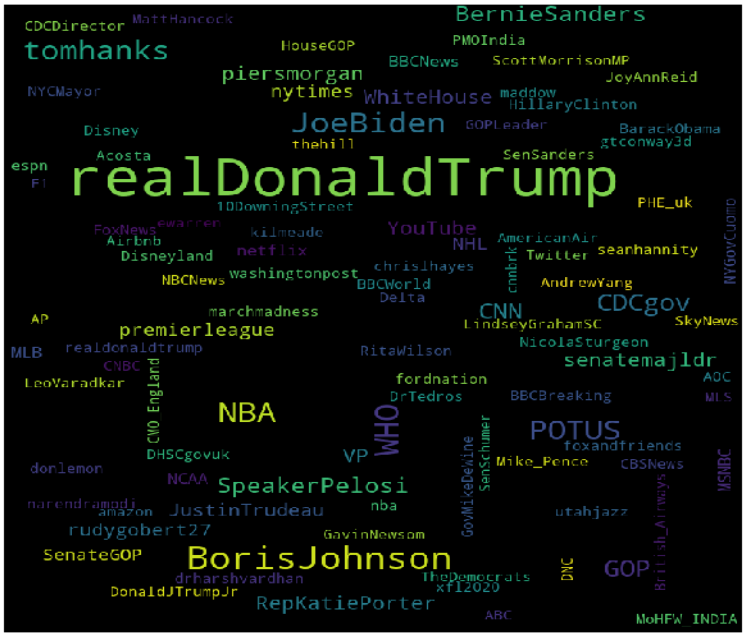
In [26]:
```
#top 10 negative sentences
english_tweets_processed.sort(desc("negative_score")).limit(10).show()
```

```
+------------------+--------------------+----------------+--------------+--------------+--------------+-----+
|         status_id|                text|   sentimentscore|negative_score|neutral_score|positive_score|label|
+------------------+--------------------+----------------+--------------+--------------+--------------+-----+
|1237907867774210048|FUCK FUCK FUCK FU...| [93.7, 6.3, 0.0]|          93.7|           6.3|           0.0|    0|
|1238244834718240771|Hell no  #Coronav...|[87.2, 12.8, 0.0]|          87.2|          12.8|           0.0|    0|
|    Italy reports 2|           016 dead"|[81.1, 18.9, 0.0]|          81.1|          18.9|           0.0|    0|
|1238093580683616256|SCARY SHIT. #Coro...|[80.5, 19.5, 0.0]|          80.5|          19.5|           0.0|    0|
|1238173590312402945|#Uncertain Strang...|[80.0, 20.0, 0.0]|          80.0|          20.0|           0.0|    0|
|1238252401875922946|Fear will kill #C...|[79.8, 20.2, 0.0]|          79.8|          20.2|           0.0|    0|
|1237992813477924864|@F1 Greedy. Shame...|[79.5, 20.5, 0.0]|          79.5|          20.5|           0.0|    0|
|1238224919835664390|#Covid_19 shit sc...|[79.3, 20.7, 0.0]|          79.3|          20.7|           0.0|    0|
|1238243761731317761|Depression #Coron...|[78.7, 21.3, 0.0]|          78.7|          21.3|           0.0|    0|
|1238225055601090563|Absolutely barbar...|[78.4, 21.6, 0.0]|          78.4|          21.6|           0.0|    0|
+------------------+--------------------+----------------+--------------+--------------+--------------+-----+
```

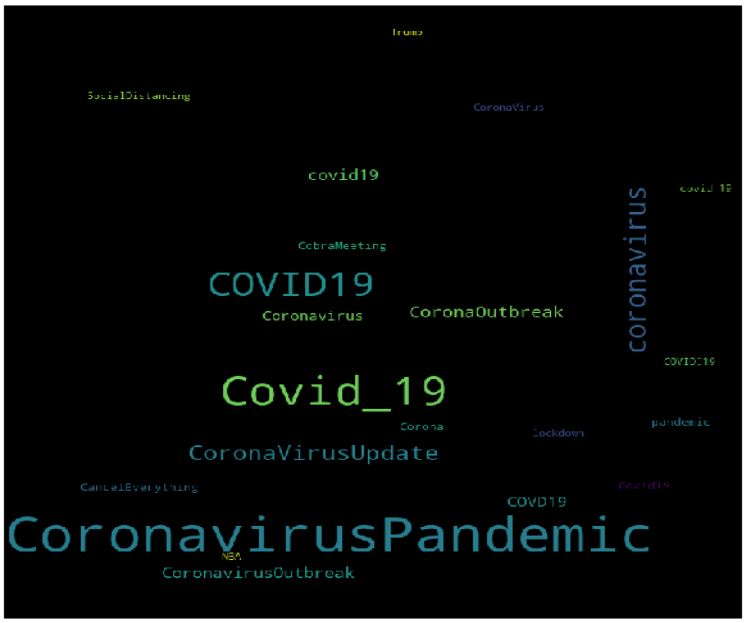
SA 3 : Frequency Analysis

- Sentiment detection is not limited detecting positives and negatives
- Public sentiment can also be captured by gauging emerging popularity of leaders and events
- Top mentioned users and hashtags in COVID-19 tweets are found and visualized using Word Cloud

SA 3: Frequency Analysis + Word Cloud



Top Users



Top Hashtags

Results and Inference

- SA1: Size of negative cluster decreased over time
- SA1: Long way to go in COVID recovery since **Negative Cluster > Positive Cluster**
- SA2: Very efficiently classified sentences for their sentiment, and thus, can be used instead of building supervised learning model from a scratch.
- SA3: Apart from regular COVID tweets, Trump and CobraMeeting were the top mentioned tags.
- SA3: Top users mentioned were American politicians and Basketball players

Future Scope

- Evaluating change in sentiment toward East Asians
- Evaluating whether top events and people were talked about in a positive way or a negative way .
- Automating the process and putting it on a website so that daily results are presented.
- Presenting region wise reports.

For code please visit: https://github.com/shrutiagarwal28/TwitterCovidSentimentAnalysis

“Truth was never told without statistics”
–Andrejs Dunkels