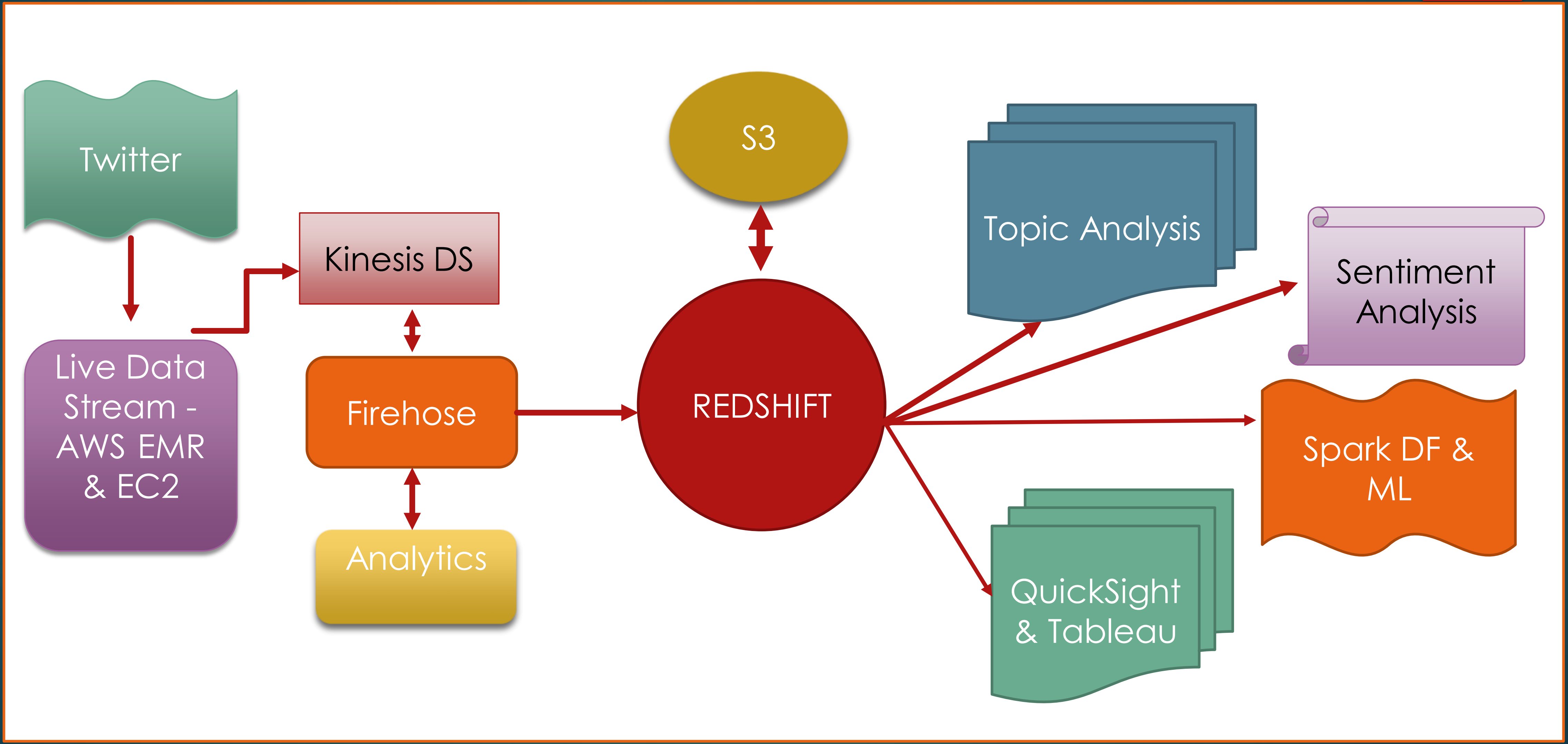
Live Streaming *to* Dashboard *to* Spark ML

- Mohammad Chowdhury

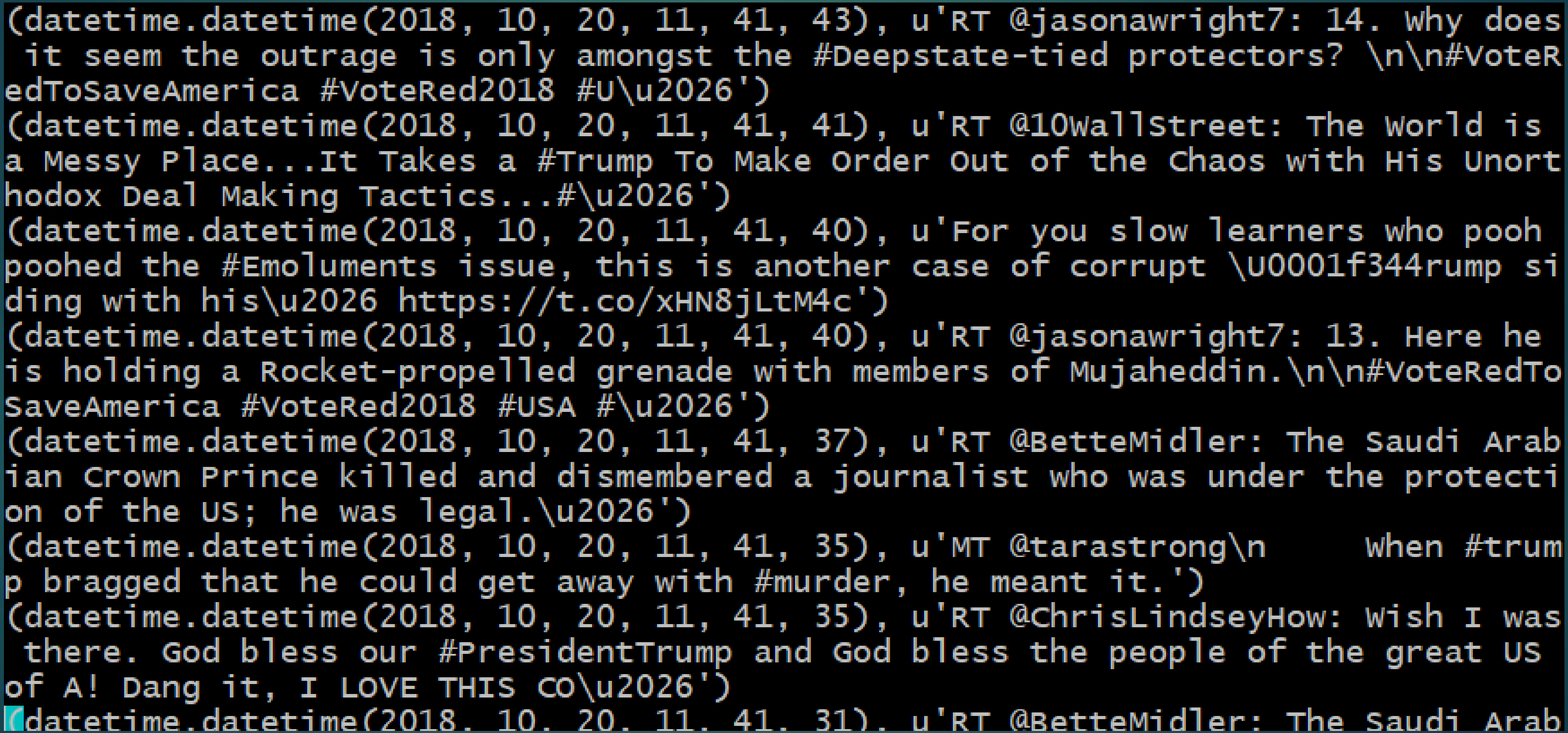
The Goal

- Build a Big Data Project using data from live twitter stream & find the sentiments & Topics using ML using Big Data technologies like
  - ❑ AWS EMR
  - ❑ AWS EC2
  - ❑ AWS S3
  - ❑ AWS Kinesis
  - ❑ AWS Redshift
  - ❑ AWS QuickSight
  - ❑ Spark
  - ❑ Tableau
  - ❑ Python
  - ❑ Linux

Data Architecture



Twitter Stream Raw Data sample



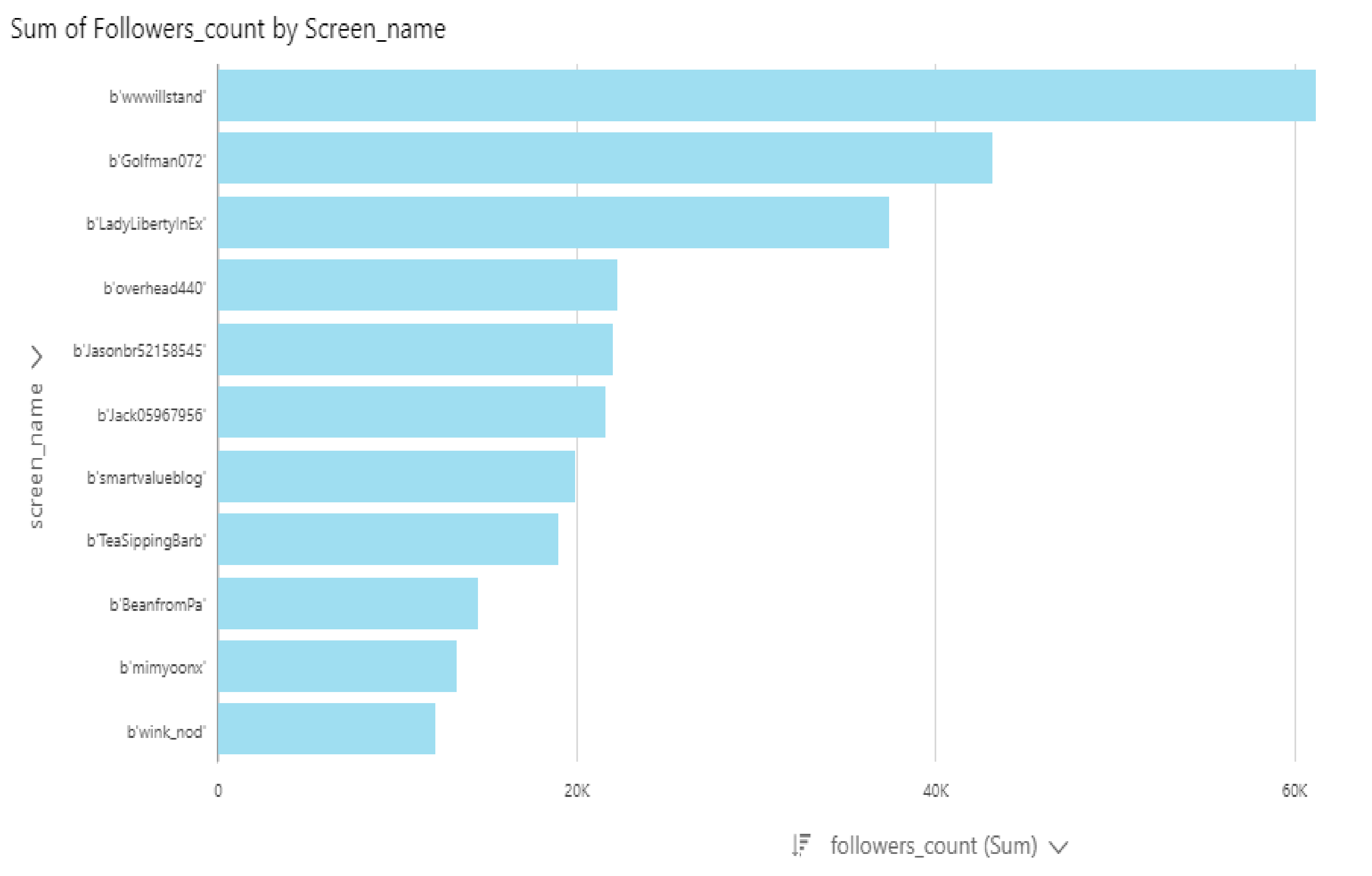
```
(datetime.datetime(2018, 10, 20, 11, 41, 43), u'RT @jasonawright7: 14. Why does
 it seem the outrage is only amongst the #Deepstate-tied protectors? \n\n#VoteR
edToSaveAmerica #VoteRed2018 #U\u2026')
(datetime.datetime(2018, 10, 20, 11, 41, 41), u'RT @10WallStreet: The World is
a Messy Place...It Takes a #Trump To Make Order Out of the Chaos with His Unort
hodox Deal Making Tactics...#\u2026')
(datetime.datetime(2018, 10, 20, 11, 41, 40), u'For you slow learners who pooh
poohed the #Emoluments issue, this is another case of corrupt \U0001f344rump si
ding with his\u2026 https://t.co/xHN8jLtM4c')
(datetime.datetime(2018, 10, 20, 11, 41, 40), u'RT @jasonawright7: 13. Here he
is holding a Rocket-propelled grenade with members of Mujaheddin.\n\n#VoteRedTo
SaveAmerica #VoteRed2018 #USA #\u2026')
(datetime.datetime(2018, 10, 20, 11, 41, 37), u'RT @BetteMidler: The Saudi Arab
ian Crown Prince killed and dismembered a journalist who was under the protecti
on of the US; he was legal.\u2026')
(datetime.datetime(2018, 10, 20, 11, 41, 35), u'MT @tarastrong\n     When #trum
p bragged that he could get away with #murder, he meant it.')
(datetime.datetime(2018, 10, 20, 11, 41, 35), u'RT @ChrisLindseyHow: Wish I was
 there. God bless our #PresidentTrump and God bless the people of the great US
of A! Dang it, I LOVE THIS CO\u2026')
(datetime.datetime(2018, 10, 20, 11, 41, 31), u'RT @BetteMidler: The Saudi Arab
```

PySpark sample code

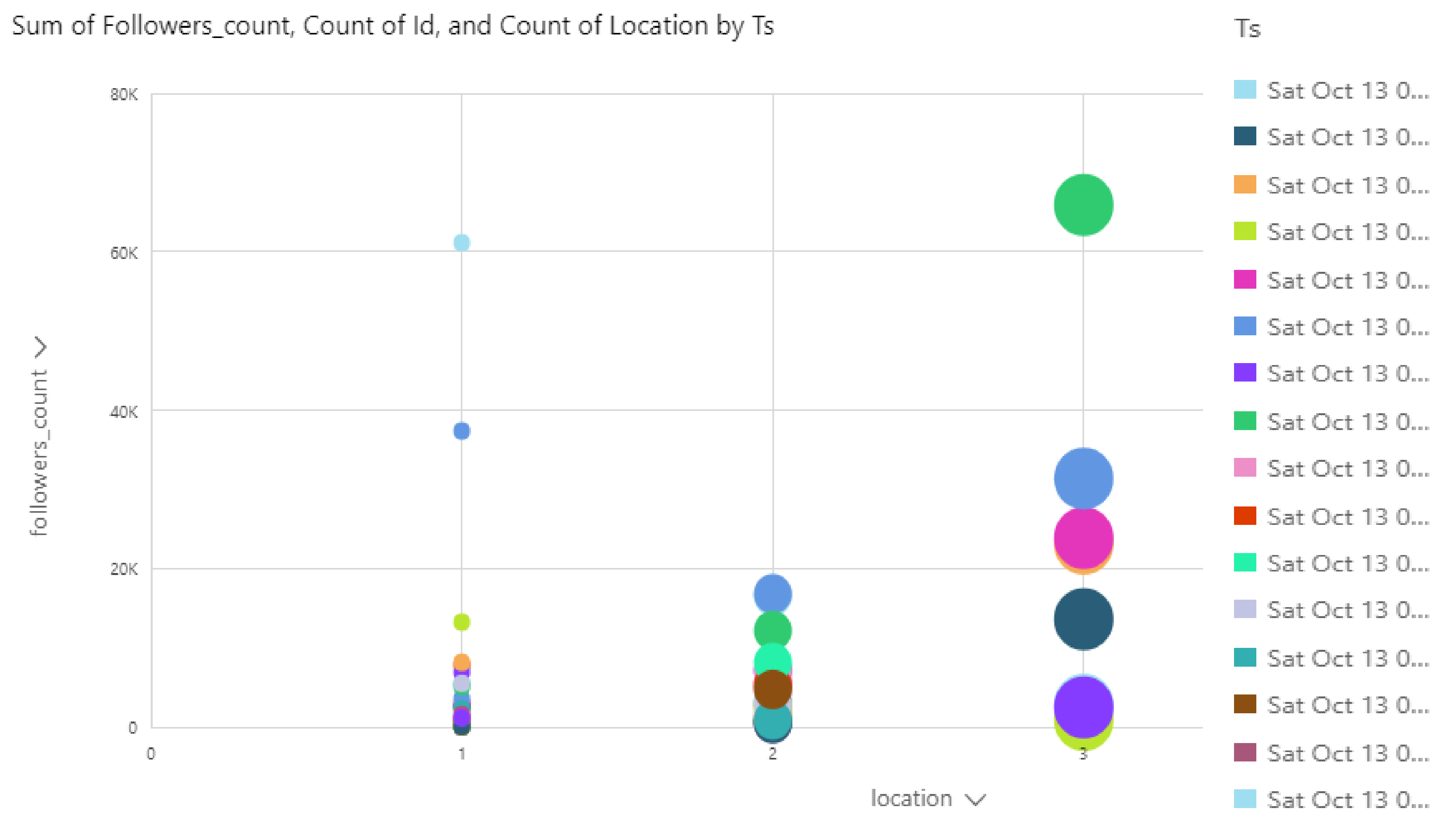
```
from pyspark.sql import SparkSession
from pyspark.sql.types import *

spark = SparkSession.builder.appName('Twitter_Aalysis').getOrCreate()
sc = spark.sparkContext
lines = sc.textFile('trump.txt')
counts = lines.flatMap(lambda line: line.split(" ")) \
            .map(lambda word: (word, 1)) \
            .sortByKey(ascending=True) \
            .reduceByKey(lambda a, b: a + b)
```

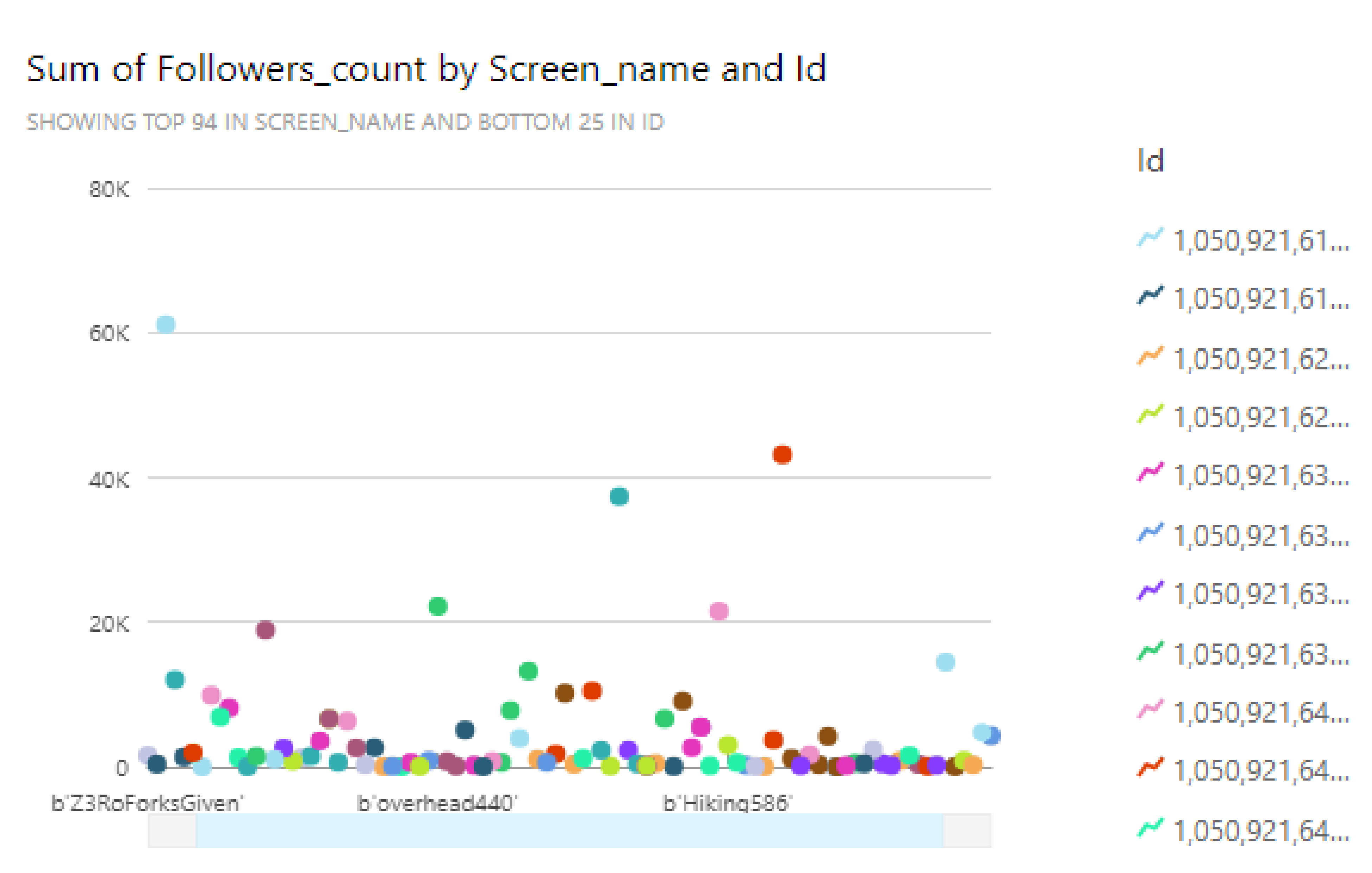
Data Analysis & Viz

Top Screen Names by Followers counts from Tweets



Location, Timestamp, followers counts from Tweets

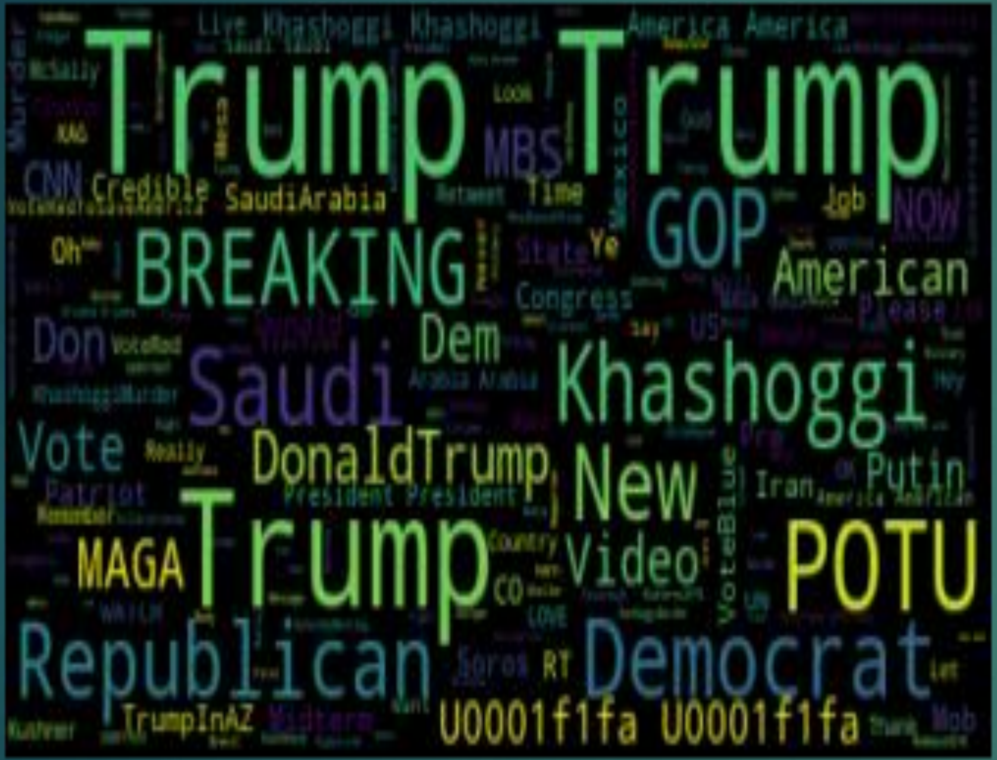


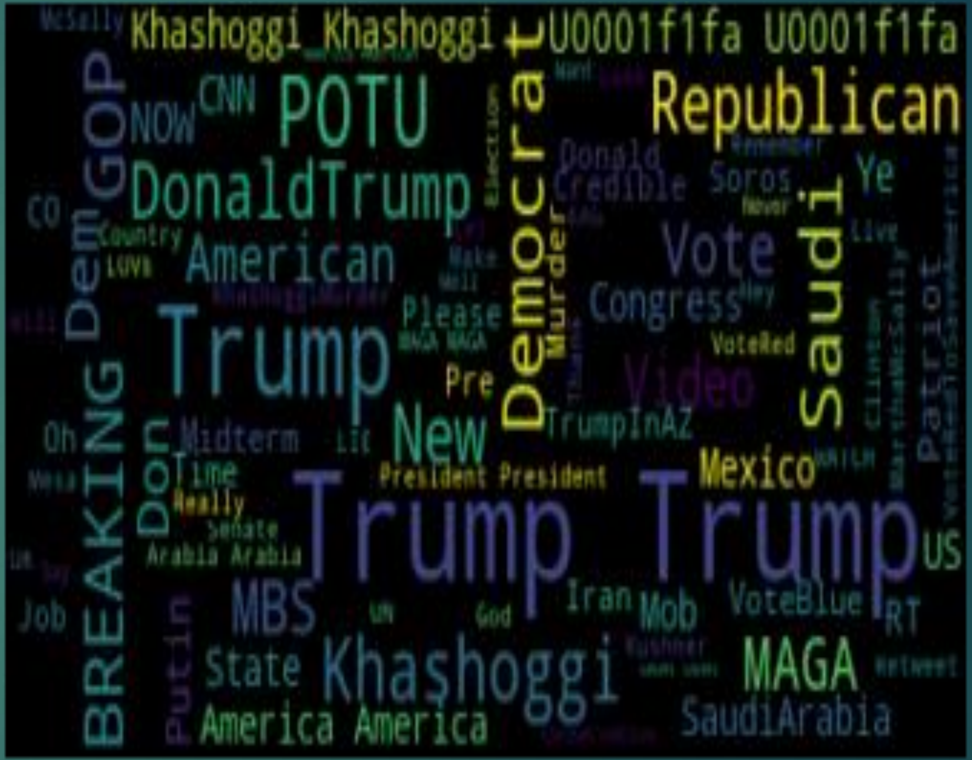


Sum of Followers_count by Screen_name and Id

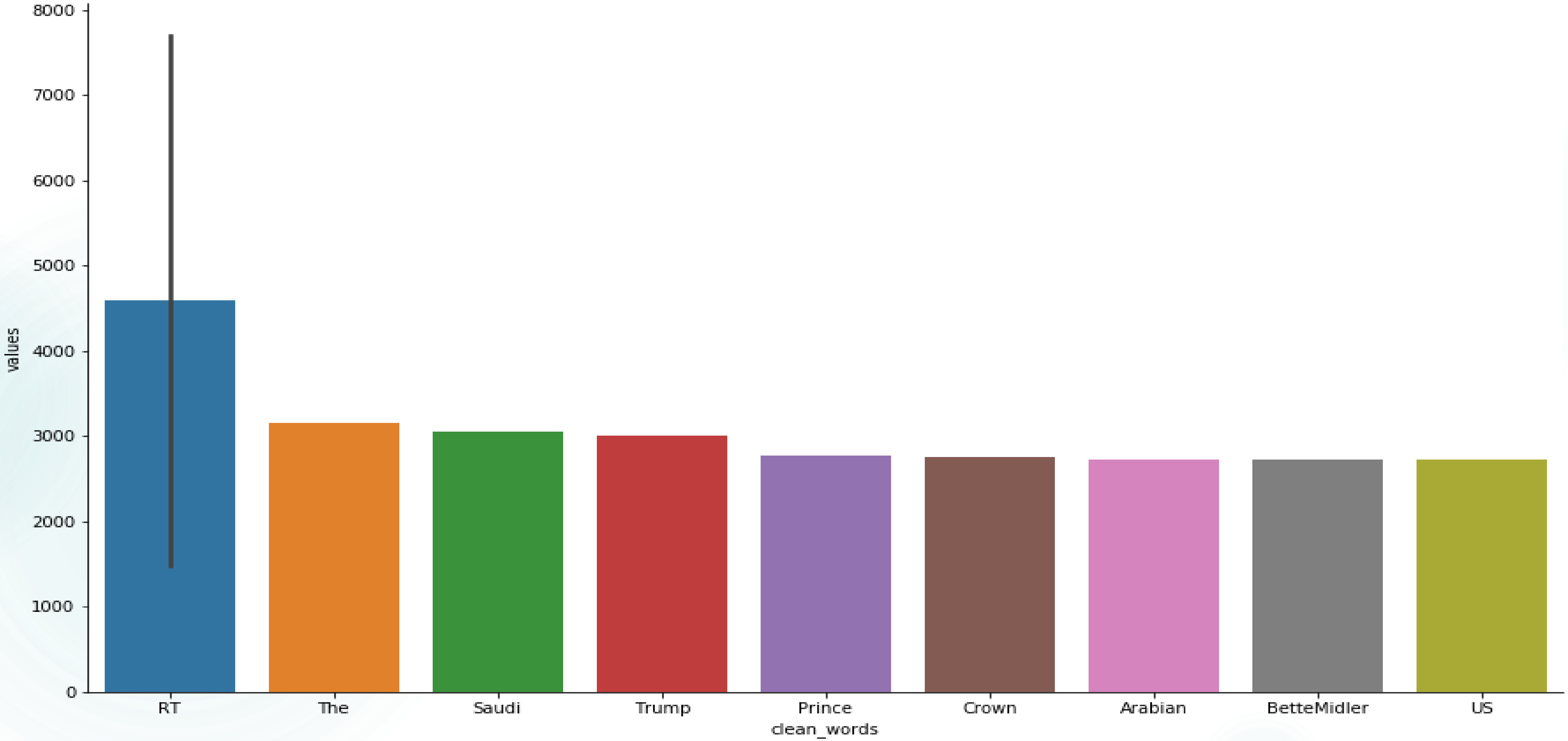
SHOWING TOP 94 IN SCREEN_NAME AND BOTTOM 25 IN ID

WordClouds


Trump Trump
Live Khashoggi Khashoggi
America America
MBS
GOP
CNN
Credible
SaudiArabia
BREAKING
American
Congress
Dem
Don
Saudi
Khashoggi
Vote
DonaldTrump
New
Putin
Iran
Patriot
MAGA
Trump
Video
POTU
VoteBlue
Republican
Democrat
Soros
RT
U0001f1fa U0001f1fa
TrumpInAZ
Midterm
Mob
Mexico
Time
Job
NOW
Oh
Ye
State
President President


McSally
Khashoggi Khashoggi
U0001f1fa U0001f1fa
CNN
NOW
POTU
GOP
Republican
Democrat
DonaldTrump
Donald
Credible
Soros
Ye
Saudi
American
Vote
Congress
Please
Murder
Dem
Trump
Video
VoteRed
Patriot
BREAKING
Don
Midterm
New
TrumpInAZ
Time
Really
Senate
President President
Mexico
Arabia Arabia
Trump Trump
US
Putin
MBS
Iran
Mob
VoteBlue
RT
Job
State
Khashoggi
MAGA
SaudiArabia
America America
Oh
CO

Top 10 Terms – PySpark Data Analysis – No ML

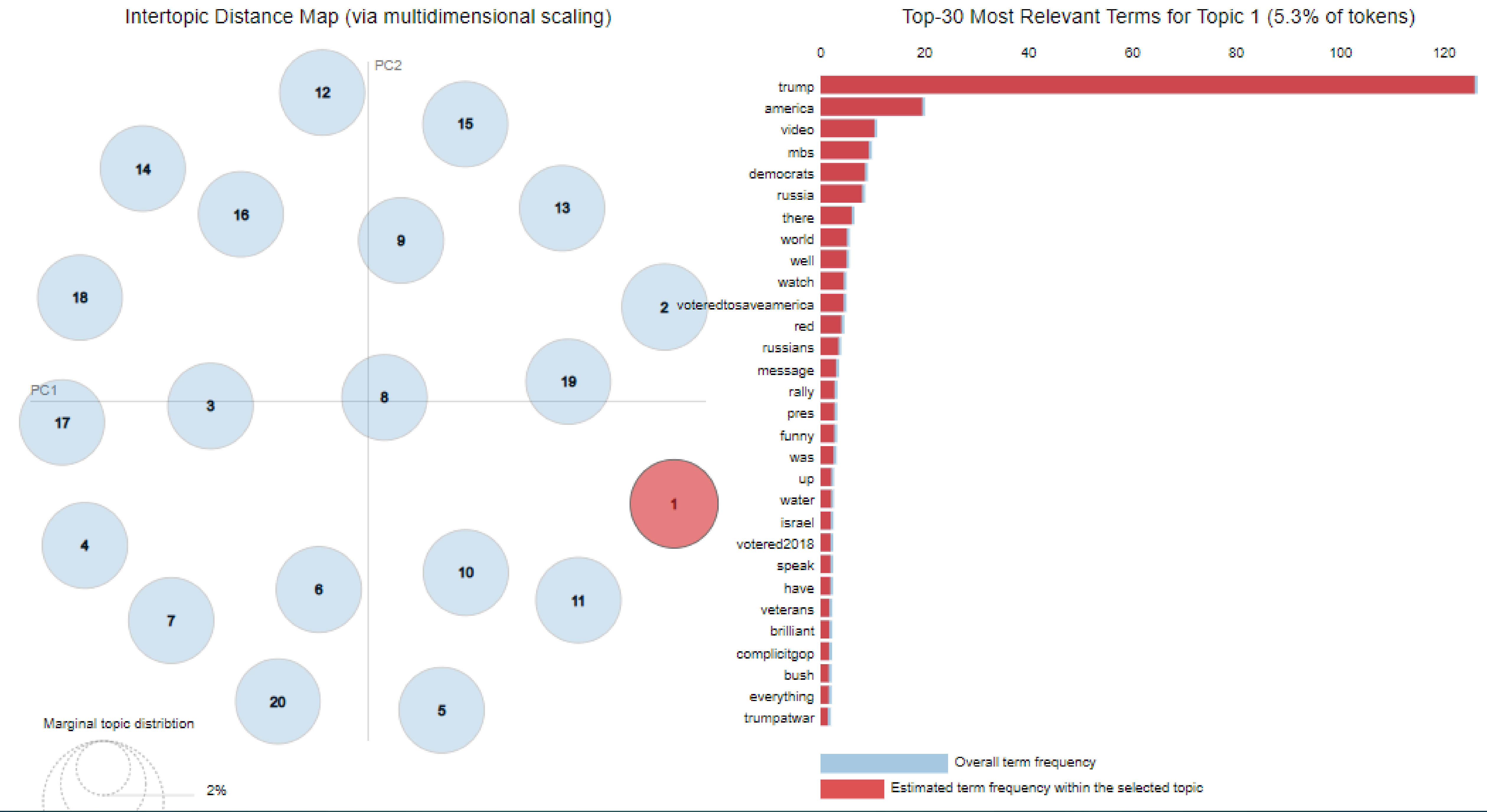


Topic Analysis | Unsupervised Learning

Latent Dirichlet Allocation (LDA)

Classical NLP TF-IDF

CountVectorizer


Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
12
15
14
16
13
9
18
2
19
3
8
17
1
4
10
11
6
7
20
5
Marginal topic distribtion
2%
Top-30 Most Relevant Terms for Topic 1 (5.3% of tokens)
0
20
40
60
80
100
120
trump
america
video
mbs
democrats
russia
there
world
well
watch
voteredtosaveamerica
red
russians
message
rally
pres
funny
was
up
water
israel
votered2018
speak
have
veterans
brilliant
complicitgop
bush
everything
trumpatwar
Overall term frequency
Estimated term frequency within the selected topic

Top trending Topics from Trump tweets

Topic 1	Topic 2	Topic 3
Trump	DonaldTrump	mega
Amercia	GOP	what
Video	American	saudiarabia
Mbs	Don	no
Democrats	who	arizona
Russia	tp	do

Sentiment Analysis results

- Positive Sentiments 66.5%
- Negative Sentiments 7.0%
- Neutral Sentiments 26.5

Thank You