

Tweet Data Pipeline

A real-time data pipeline for querying and visualizing tweets



Presented by

Devashish Kulkarni
Mili Gera
Heesuk Jang
Matt Whittaker

W205 Data Engineering
MIDS, UC Berkeley
Fall 2021

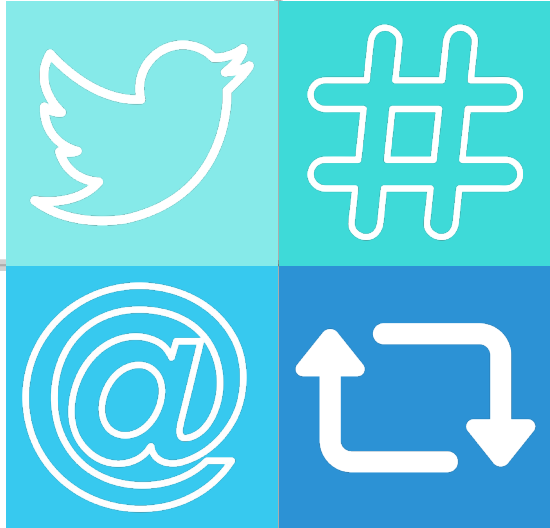


OVERVIEW

Problem: Twitter has become a prominent source of real-time content, yet lacks the ability to perform Tweet analytics beyond the limited web user interface



Solution: Create a data pipeline to gather tweets, fully customizable based on filtering rules and criteria

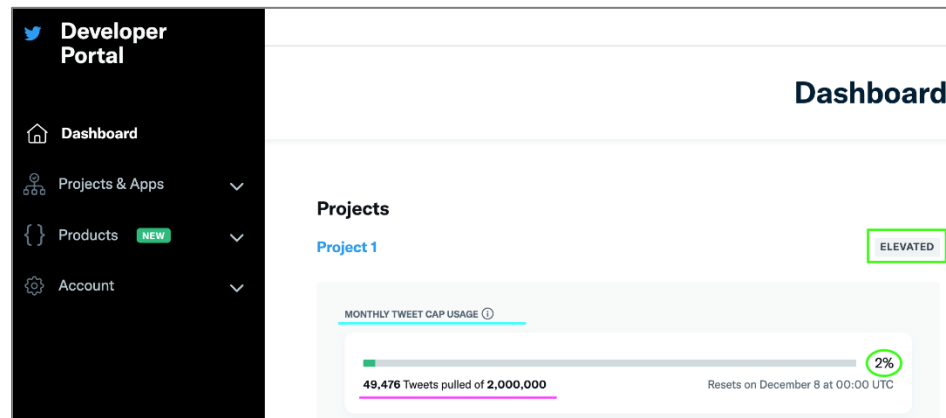
On-demand Data available on-demand from the Twitter API using python, with messages consumed and produced through Kafka and Spark		Customizable Tweets can be aggregated by hashtag, location or topic
Analytics Robust data processing allowing for SQL to be used for advanced analytics and querying		Structured Data Tweets are compiled into standardized format for aggregation into logical dataframe based structure

Application: Customized filtering rules and criteria for Climate Change topics on Twitter

DATA PIPELINE



STEP 1: Twitter Developer Account Setup



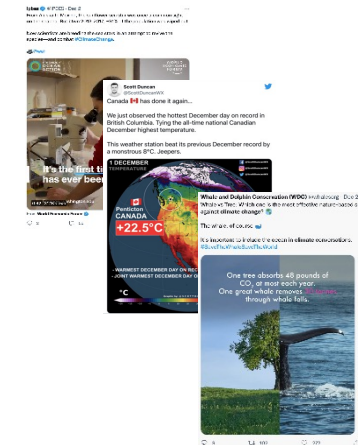
STEP 2: Filtered Stream Endpoints

POST /2/tweets/search/stream/rules

GET /2/tweets/search/stream



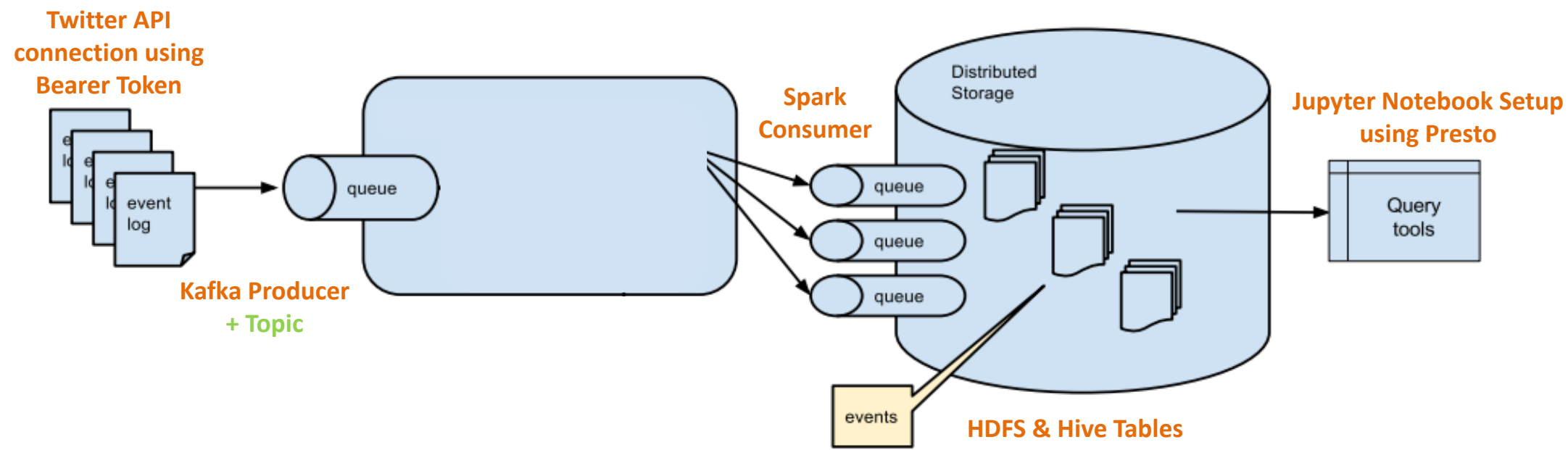
STEP 3: Read Tweets



STEP 4: Tweets Ingestion and Loading in relation to Kafka Topic



STEP 5: ETL, Data Storage, and Data Querying



TOOLS

Tweet streaming (via Twitter API)  tools incorporated in Google Cloud Platform (GCP) and Jupyter

D

Docker Compose

Used to create containers of various services required to construct the data pipeline

Z

Zookeeper

Used to track the status of nodes in the Kafka cluster and maintain a list of kafka topics and messages.

K

Kafka

Used as a broker to receive messages produced by Twitter and store them in a topic

S

Spark

Used to consume messages stored in a Kafka topic, perform the ETL transformations and store them into HDFS as parquet files

H

HDFS

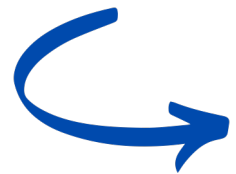
Distributed files system from Hadoop that is used to store the tweet streaming data in tables

P

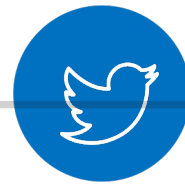
Presto / Hive

Used to query the tables stored in HDFS

Twitter
API

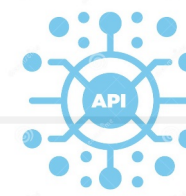


FILTERED STREAM REQUEST



Tweet object

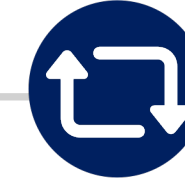
- Basic Building Block of all things Twitter
- Root level 'fields'
 - id
 - text
 - created_at
- Parent to children objects
 - user
 - media
 - place



Twitter API rules

Rules used for this analysis:

- Tweets from leading climate change activist organizations
- Includes hashtags or key phrases linked to climate change
- Accounts that mention (@) a leading organization



Requested fields

Requested fields in this analysis:

- Tweet fields
 - Root level fields
 - Metrics (likes, retweets, replies)
- User fields
 - Name
 - Username
 - Location
- Place fields
 - Country
 - Country code

+ BEARER_TOKEN

Streaming Data

```

{
  "data": {
    "author_id": "72393875",
    "geo": {},
    "id": "1464585858544197636",
    "text": "@smolrobots No Nazguls around though. (going on her film version)"
  },
  "includes": {
    "users": [
      {
        "id": "72393875",
        "location": "UK",
        "name": "James Green #Rejoin",
        "username": "Jim1810"
      },
      {
        "id": "933815913769512961",
        "location": "small robot development lab",
        "name": "small robots",
        "username": "smolrobots"
      }
    ]
  },
  "matching_rules": [
    {
      "id": "1464585887493148685",
      "tag": "climate change text only"
    }
  ]
}

```

Tweet



small robots @smolrobots · Nov 27
 ...

You know climate change is bad when we run out of human names for storms and move onto the elvish ones.

8
 35
 198



James Green #Rejoin
...

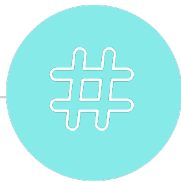
@Jim1810

Replying to [@smolrobots](#)

No Nazguls around though. (going on her film version)

5:24 AM · Nov 27, 2021 · Twitter for Android

FINAL TABLE TO BE QUERIED using PRESTO



Selected Tweet Metadata

- Author ID • Tweet ID • Tweet Created At • Retweet Count • Reply Count • Like Count • User Location • User Fullname • Username • Tweet Text • Hashtag Name • URL



- To HDFS through Data pipeline



- Converted to external table by Hive



- Queried by Presto and loaded in pd.DataFrame

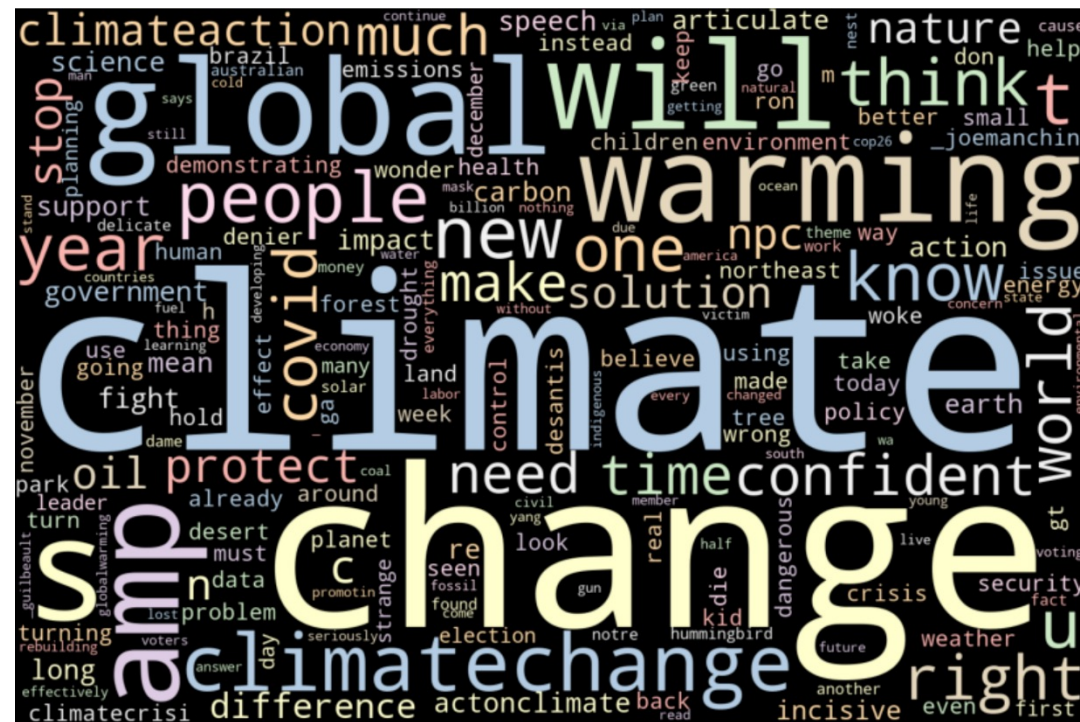


author_id	tweet_id	tweet_created_at	retweet_count	reply_count	like_count	user_location	user_fullname	username	tweet_text	hashtag_name	url
1466505129176322051	1729193269	2021-12-02T20:30:43.000Z	1	0	0	None	JFSebastian146	JFSebastian146	RT @shabazzshareef: Aerial view of flooded nei...	Houston, HurricaneHarvey, climatechange, dr	[https://t.co/RnLQrSKE3w]
1466497205255032839	36165350	2021-12-02T19:59:14.000Z	25	0	0	Portland, OR	Bob Leonard	BobOne4All	RT @AaravSeth_: #HeavyRains & hail fall ag...	HeavyRains, flooding, Oman	[]
1466504728699981829	2192416092	2021-12-02T20:29:08.000Z	8	0	0	None	DrGem	DrGem2015	RT @DrGem2015: Why global warming is more bene...	Greta, globalwarming, climatechan	[https://t.co/AqTI0blcO4]
1466503856574062599	1328069407164215303	2021-12-02T20:25:40.000Z	9	0	0	None	Independent Scientists for Nature	IndependentSci4	RT @biofuelwatch: There is way too much #Green...	Greenwashing, aviationfest	[]
1466500101036195843	2475682276	2021-12-02T20:10:44.000Z	2	0	0	None	Source_BTC	Source_33_	RT @Tradinator33: #GlobalWarming \n`\"(ツ)\"_\" h...	GlobalWarming	[https://t.co/tA8PUFJ74v]

PIPELINE DEMO

ANALYTICS with VISUALIZATIONS

Q1. The Most Trending Words in the Entire Tweets



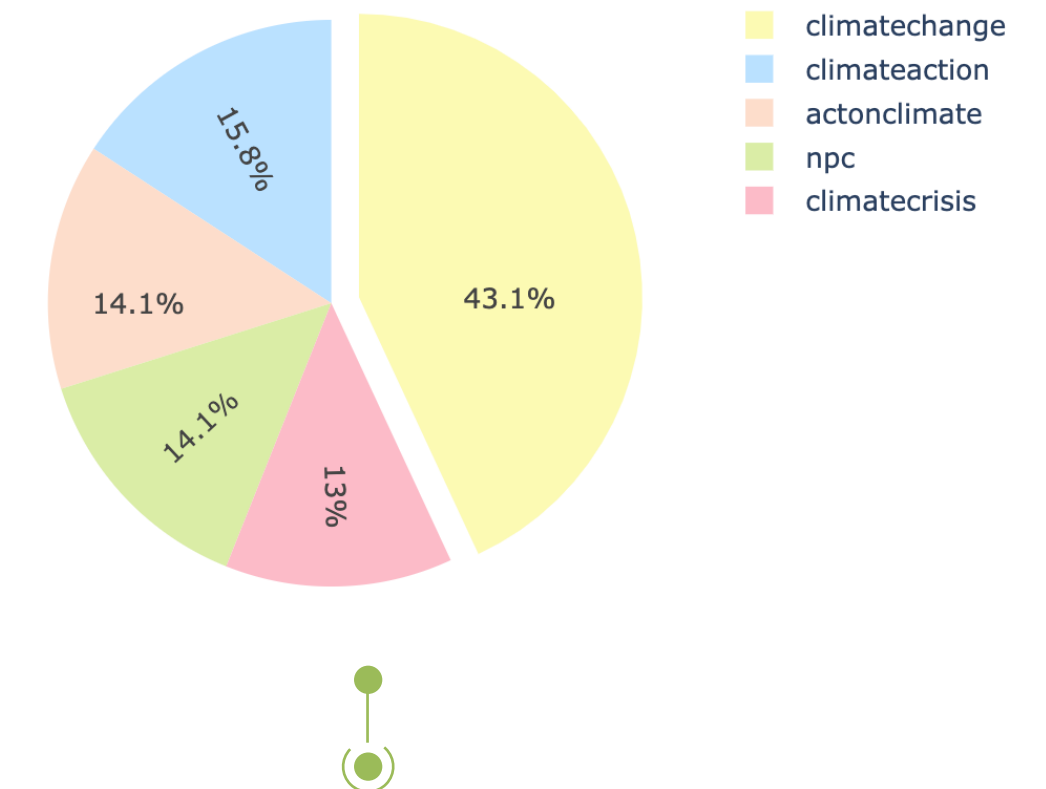
- Climate
- Change
- Global
- Warming
- World

Q2. The Most Trending Words in the Top 100 Tweets with the Most Retweets



- Oil
- Gas
- Environmentalism
- Greenland
- exploration

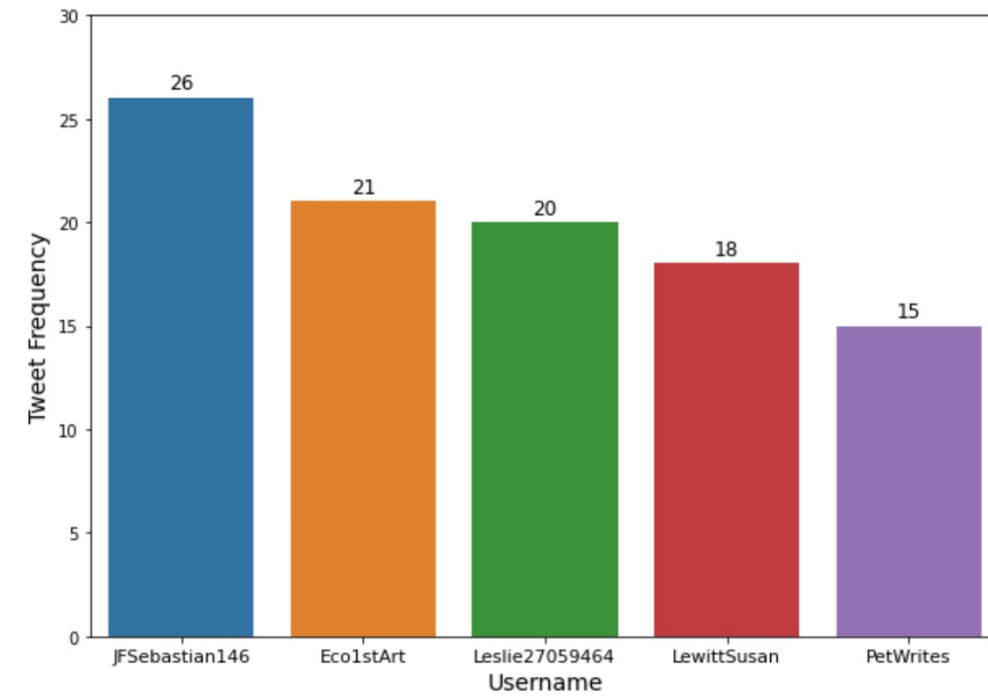
Q3. Top 5 Most Trending Hashtags



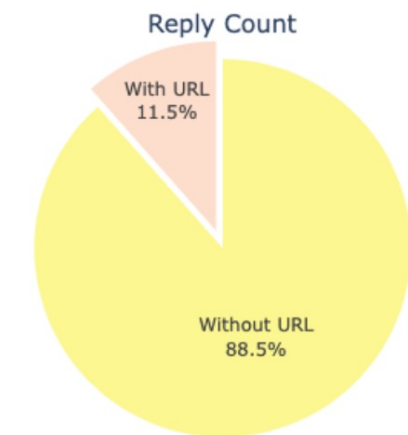
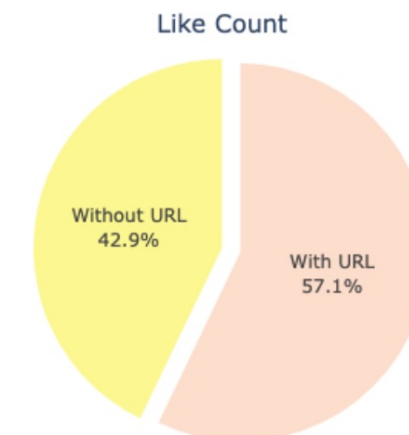
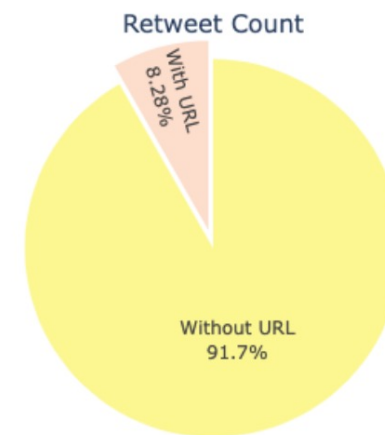
- #ClimateChange
- #ClimateAction
- #ActionClimate
- #NPC
- #ClimateCrisis

ANALYTICS with VISUALIZATIONS

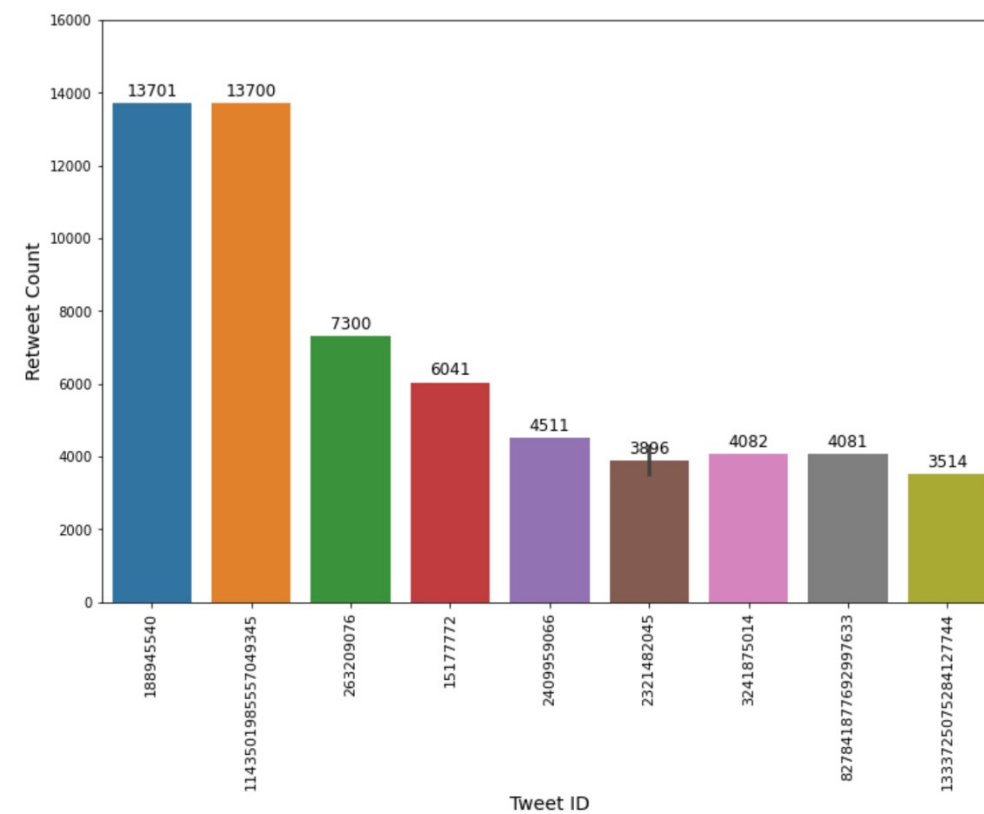
Q4. Top 5 Most Active Tweeters



Q6. Count of Retweets, Likes, and Replies Having Links



Q5. Top 10 Tweets with the Most Retweets



Q7. Count of Retweets and Likes Having Hashtags

