CS 4371.501 Introduction to Big Data Management and Analytics (Fall 2021)

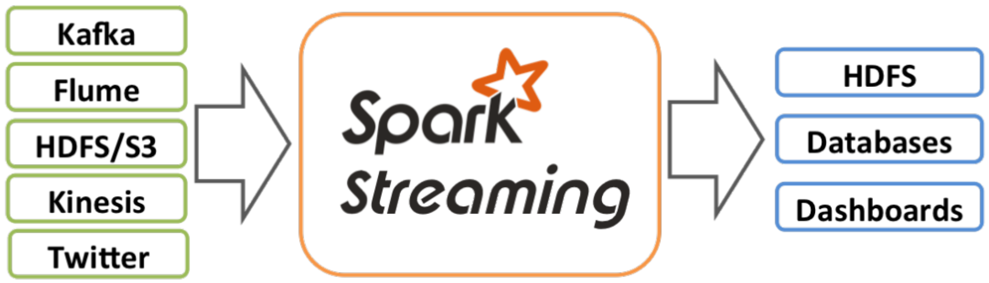
Project: Spark Streaming Background

This project deals with real-time streaming data arriving from twitter streams.

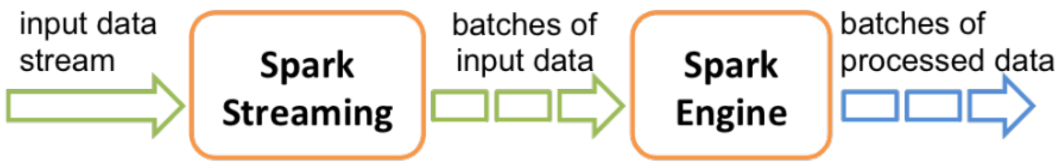
1. Spark Streaming

Spark Streaming is an extension of the core Spark API that enables scalable, high-throughput, fault-tolerant stream processing of live data streams. Data can be ingested from many sources like Kafka, Flume, Kinesis, or TCP sockets, and can be processed using complex algorithms expressed with high-level functions like *map*, *reduce*, *join* and *window*.

Finally, processed data can be pushed out to filesystems, databases, and live dashboards. In fact, you can apply Spark's **machine learning** and **graph processing** algorithms on data streams.



Internally, it works as follows. Spark Streaming receives live input data streams and divides the data into batches, which are then processed by the Spark engine to generate the final stream of results in batches.



Spark Streaming provides a high-level abstraction called discretized stream or DStream, which represents a continuous stream of data. DStreams can be created either from input data streams from sources such as Kafka, Flume, and Kinesis, or by applying high-level operations on other DStreams. Internally, a DStream is represented as a sequence of RDDs.

Please checkout this website for details.

2. Source

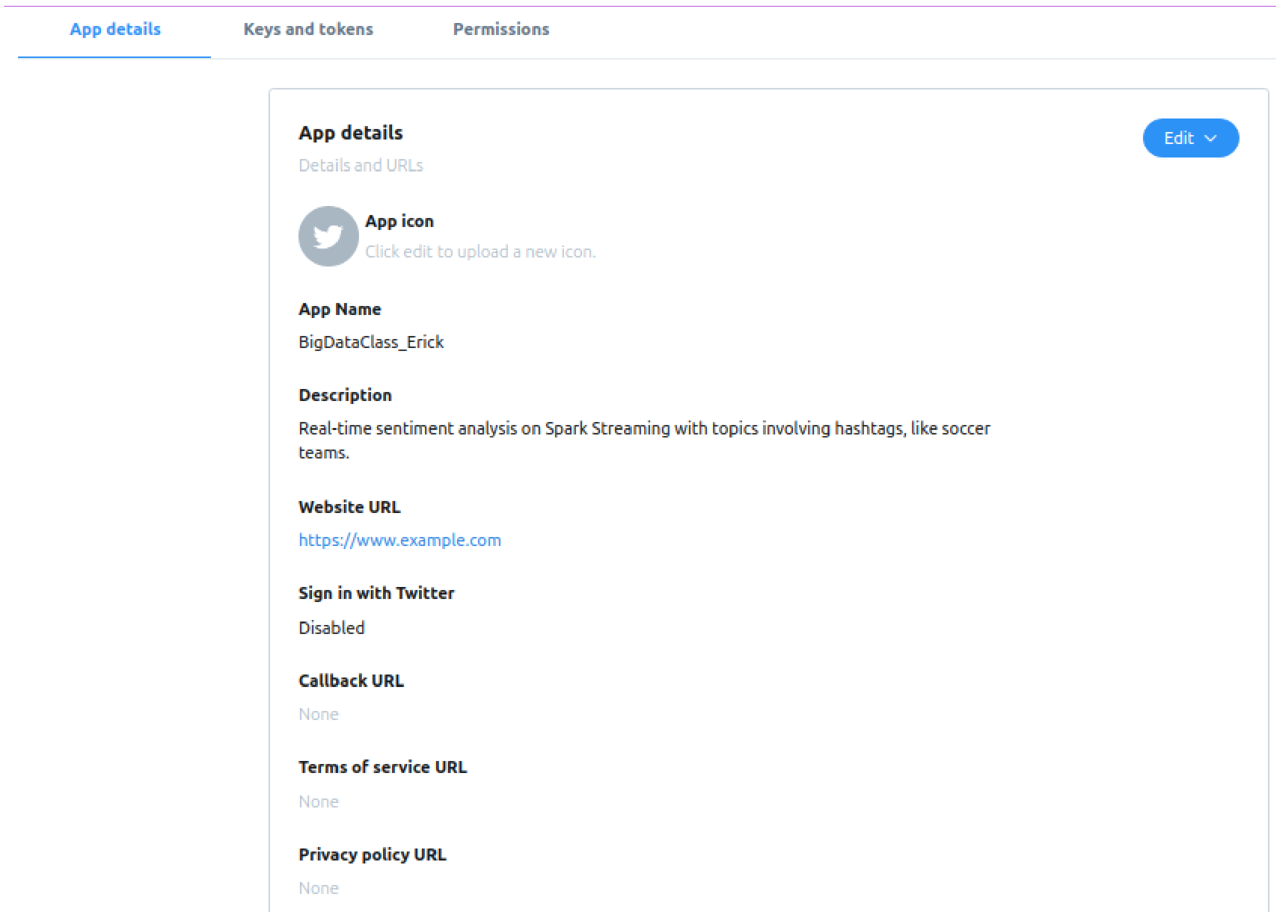
In this assignment, we will use **twitter** as our stream source. To access twitter stream, please see the following instructions:

2.1. First create your own account at https://apps.twitter.com/

2.2. Click on ***Create an App*** to create your own Twitter app.

2.3. Fill out the form and submit

2.4. Once you have created your twitter app, you will observe a form similar to this:



App details
Details and URLs

App icon
Click edit to upload a new icon.

App Name
BigDataClass_Erick

Description
Real-time sentiment analysis on Spark Streaming with topics involving hashtags, like soccer teams.

Website URL
https://www.example.com

Sign in with Twitter
Disabled

Callback URL
None

Terms of service URL
None

Privacy policy URL
None

2.5. Click on **Keys and tokens**, generate your keys (you will need them to add in stream.py code)

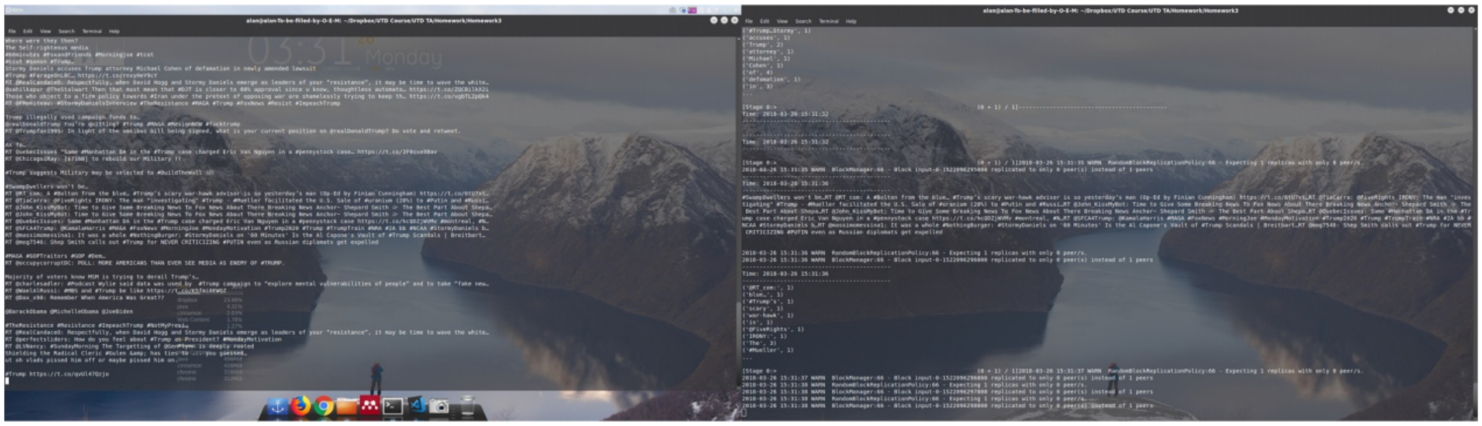
> Consumer Key (API key)
> Consumer Secret (API Key secret)
> Access Token
> Access Token Secret

2.6. Crawling and Create a Stream

> **Crawling** – refer to *stream.py* for details
> **Create a Stream** – refer to *spark.py* for details

This setup code related to stream.py and spark.py are available on elearning website.

Once you run these codes (run stream.py first and then run spark.py), you are going to observe such a window



```
Where were they then?
The Self-righteous media
#60minutes #FoxandFriends #Morningjoe #tcot
#ccot #qanon #Trump…
Stormy Daniels accuses Trump attorney Michael Cohen of defamation in newly amended lawsuit
#Trump #FarageOnLBC… https://t.co/rnwyHerNcY
RT @RealCandaceO: Respectfully, when David Hogg and Stormy Daniels emerge as leaders of your "resistance", it may be time to wave the white…
@sahilkapur @TheStalwart Then that must mean that #DJT is closer to 60% approval since u know, thoughtless automato… https://t.co/ZQCBilkX21
Those who object to a firm policy towards #Iran under the pretext of opposing war are shamelessly trying to keep th… https://t.co/vgbTL2pQk4
RT @FMonsteau: #StormyDanielsInterview #TheResistance #MAGA #Trump #FoxNews #Resist #ImpeachTrump

Trump illegally used campaign funds to…
@realDonaldTrump You're quitting? #Trump #MAGA #ResignNOW #fucktrump
RT @Trumpfan1995: In light of the omnibus bill being signed, what is your current position on @realDonaldTrump? Do vote and retweet.

As to…
RT QuebecIssues "Same #Manhattan DA in the #Trump case charged Eric Van Nguyen in a #pennystock case… https://t.co/2P8sxe8Bav
RT @ChicagoIRay: [$716B] to rebuild our Military !!

#Trump suggests Military may be selected to #BuildTheWall

#SwampDwellers won't be…
RT @RT_com: A #Bolton from the blue… #Trump's scary war-hawk adviser is so yesterday's man (Op-Ed by Finian Cunningham) https://t.co/6tU7xt…
RT @TiaCarra: @FiveRights IRONY: The man "investigating" #Trump - #Mueller facilitated the U.S. Sale of #uranium (20%) to #Putin and #Russi…
RT @John_KissMyBot: Time to Give Some Breaking News To Fox News About There Breaking News Anchor~ Shepard Smith -> The Best Part About Shepa…
RT @John_KissMyBot: Time to Give Some Breaking News To Fox News About There Breaking News Anchor~ Shepard Smith -> The Best Part About Shepa…
RT @QuebecIssues: Same #Manhattan DA in the #Trump case charged Eric Van Nguyen in a #pennystock case https://t.co/hcQDZjWVMv #montreal, #N…
RT @SFCA4Trump: @KamalaHarris #MAGA #FoxNews #MorningJoe #MondayMotivation #Trump2020 #Trump #TrumpTrain #NRA #2A bb #NCAA #StormyDaniels b…
RT @massimomessinal: It was a whole #NothingBurger: #StormyDaniels on '60 Minutes' Is the Al Capone's Vault of #Trump Scandals | Breitbart…
RT @mog7546: Shep Smith calls out #Trump for NEVER CRITICIZING #PUTIN even as Russian diplomats get expelled

#MAGA #GOPTraitors #GOP #Dem…
RT @occupycorruptDC: POLL: MORE AMERICANS THAN EVER SEE MEDIA AS ENEMY OF #TRUMP.

Majority of voters know MSM is trying to derail Trump's…
RT @charlesadler: #Podcast Wylie said data was used by  #Trump campaign to "explore mental vulnerabilities of people" and to take "fake new…
RT @WaelAlRussi: #MBS and #Trump be like https://t.co/K5fWi8EWQZ
RT @Dax_x98: Remember When America Was Great??

@BarackObama @MichelleObama @JoeBiden

#TheResistance #Resistance #ImpeachTrump #NotMyPresi…
RT @RealCandaceO: Respectfully, when David Hogg and Stormy Daniels emerge as leaders of your "resistance", it may be time to wave the white…
RT @perfectsliders: How do you feel about #Trump as President? #MondayMotivation
RT @LVNancy: #SundayMorning The Targetting of @GenFlynn is deeply rooted
Shielding the Radical Cleric #Gulen &amp; has ties to you guessed…
ut oh vlads pissed him off or maybe pissed him on.

#Trump https://t.co/qvUl47Qzjo
```

```
('#Trump…Stormy', 1)
('accuses', 1)
('Trump', 2)
('attorney', 1)
('Michael', 1)
('Cohen', 1)
('of', 4)
('defamation', 1)
('in', 3)
...

[Stage 0:>                                                          (0 + 1) / 1]------------------------------------------
Time: 2018-03-26 15:31:32
-------------------------------------------

-------------------------------------------
Time: 2018-03-26 15:31:32
-------------------------------------------

[Stage 0:>                                                          (0 + 1) / 1]2018-03-26 15:31:35 WARN  RandomBlockReplicationPolicy:66 - Expecting 1 replicas with only 0 peer/s.
2018-03-26 15:31:35 WARN  BlockManager:66 - Block input-0-1522096295000 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 2018-03-26 15:31:36
-------------------------------------------
#SwampDwellers won't be…RT @RT_com: A #Bolton from the blue… #Trump's scary war-hawk adviser is so yesterday's man (Op-Ed by Finian Cunningham) https://t.co/6tU7xt…RT @TiaCarra: @FiveRights IRONY: The man "inves
tigating" #Trump - #Mueller facilitated the U.S. Sale of #uranium (20%) to #Putin and #Russi…RT @John_KissMyBot: Time to Give Some Breaking News To Fox News About There Breaking News Anchor~ Shepard Smith -> The
 Best Part About Shepa…RT @John_KissMyBot: Time to Give Some Breaking News To Fox News About There Breaking News Anchor~ Shepard Smith -> The Best Part About Shepa…RT @QuebecIssues: Same #Manhattan DA in the #Tr
ump case charged Eric Van Nguyen in a #pennystock case https://t.co/hcQDZjWVMv #montreal, #N…RT @SFCA4Trump: @KamalaHarris #MAGA #FoxNews #MorningJoe #MondayMotivation #Trump2020 #Trump #TrumpTrain #NRA #2A bb #
NCAA #StormyDaniels b…RT @massimomessinal: It was a whole #NothingBurger: #StormyDaniels on '60 Minutes' Is the Al Capone's Vault of #Trump Scandals | Breitbart…RT @mog7546: Shep Smith calls out #Trump for NEVER
 CRITICIZING #PUTIN even as Russian diplomats get expelled


2018-03-26 15:31:36 WARN  RandomBlockReplicationPolicy:66 - Expecting 1 replicas with only 0 peer/s.
2018-03-26 15:31:36 WARN  BlockManager:66 - Block input-0-1522096296000 replicated to only 0 peer(s) instead of 1 peers
-------------------------------------------
Time: 2018-03-26 15:31:36
-------------------------------------------
('@RT_com:', 1)
('blue…', 1)
('#Trump's', 1)
('scary', 1)
('war-hawk', 1)
('is', 1)
('@FiveRights', 1)
('IRONY:', 1)
('The', 3)
('#Mueller', 1)
...

[Stage 0:>                                                          (0 + 1) / 1]2018-03-26 15:31:37 WARN  RandomBlockReplicationPolicy:66 - Expecting 1 replicas with only 0 peer/s.
2018-03-26 15:31:37 WARN  BlockManager:66 - Block input-0-1522096296800 replicated to only 0 peer(s) instead of 1 peers
2018-03-26 15:31:38 WARN  RandomBlockReplicationPolicy:66 - Expecting 1 replicas with only 0 peer/s.
2018-03-26 15:31:38 WARN  BlockManager:66 - Block input-0-1522096297800 replicated to only 0 peer(s) instead of 1 peers
2018-03-26 15:31:38 WARN  RandomBlockReplicationPolicy:66 - Expecting 1 replicas with only 0 peer/s.
2018-03-26 15:31:38 WARN  BlockManager:66 - Block input-0-1522096298000 replicated to only 0 peer(s) instead of 1 peers
```

3. Task

You are asked to implement a framework using Apache Spark Streaming to perform sentiment analysis of particular hashtags in twitter data in real time. For example, suppose we want to run the sentiment analysis (e.g.,positive, neutral, negative) for all the tweets containing a given hashtag, say "*#staysafe*", and later show some statistics.

The first step for this task is to get the tweets via scrapper (stream.py). Next, you will apply any off-the-shelf pre-trained model for sentiment analysis over the collected tweet messages. Note that gathering data related to location (where the tweet was generated) may also be useful for analyzing later (Don't worry, stream.py already gets the data for you). Finally, you will visualize your findings using either Spark SQL or ElasticSearch/Kibana (preferable!).

4. Putting all together

The flow for the entire project will be:

Scrapper ---> Sentiment Analyzer ---> Visualizing (Spark SQL or ElasticSearch/Kibana)

As aforementioned, the skeleton of the framework is provided in python on eLearning. The parts you will need to change are marked as comments. Always remember that this is YOUR project, and you are free to change 100% of the codes, including the provided skeleton.

4.1. Scrapper: We provide a sample scrapper (stream.py). However, you need to extend the code to support the following functionality described below.

The scrapper collects tweets and pre-process them for analytics. It is a standalone program written in Python and should perform the following:

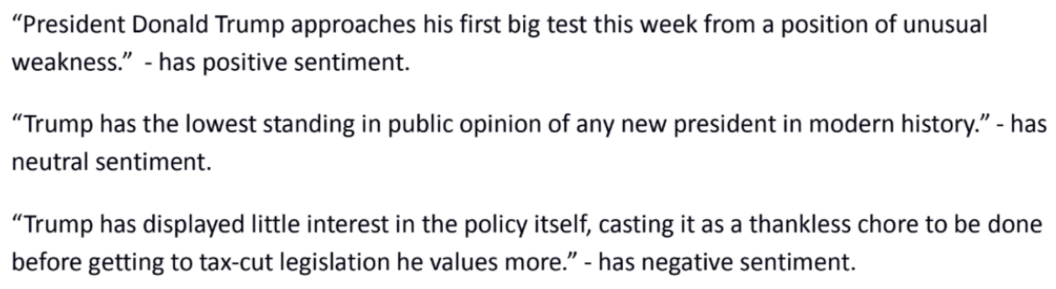
(i) Collect tweets in real-time with a particular hashtag. For example, we will collect all tweets with #*staysafe* or ***any other you prefer***.

(ii) After getting tweets, we pre-process them by removing emoji symbols and special characters, discarding any noise on tweets. Note that the returned tweet contains both the metadata (e.g., location) and text contents. You will have to keep at least the text content and the location metadata.

(iii) After pre-processing, you need to convert the location metadata of each tweet back to its geolocation info by calling either google geo API or geopy (or any other you prefer) and send the text and geolocation info to spark streaming. Note that this step can also be performed on the spark.py side.

(iv) Your scrapper program should run infinitely and should take hashtags as input parameters while running.

4.2. Sentiment Analyzer: Sentiment Analyzer determines whether a piece of tweet is positive, neutral or negative. For example,



"President Donald Trump approaches his first big test this week from a position of unusual weakness." - has positive sentiment.

"Trump has the lowest standing in public opinion of any new president in modern history." - has neutral sentiment.

"Trump has displayed little interest in the policy itself, casting it as a thankless chore to be done before getting to tax-cut legislation he values more." - has negative sentiment.

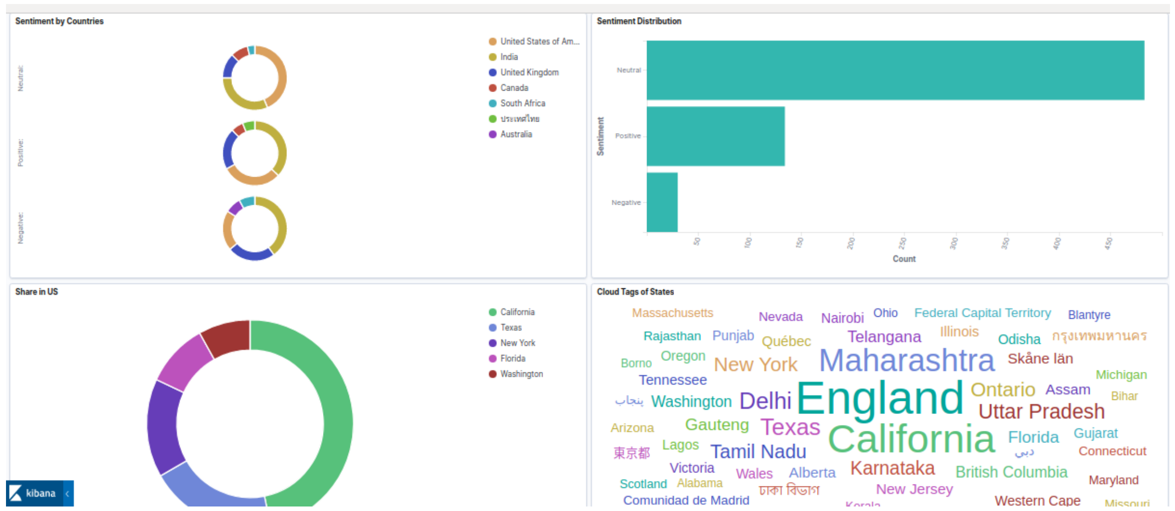
You can use any third-party sentiment analyzer like Stanford CoreNLP, NLTK or any other software or pre-trained model for sentiment analysis.

In summary, for each hashtag, you perform sentiment analysis using sentiment analysis tools discussed above and output sentiment and geolocation of each tweet to some external bases (either save in a json file or send them to elasticsearch index for visualization on Kibana).

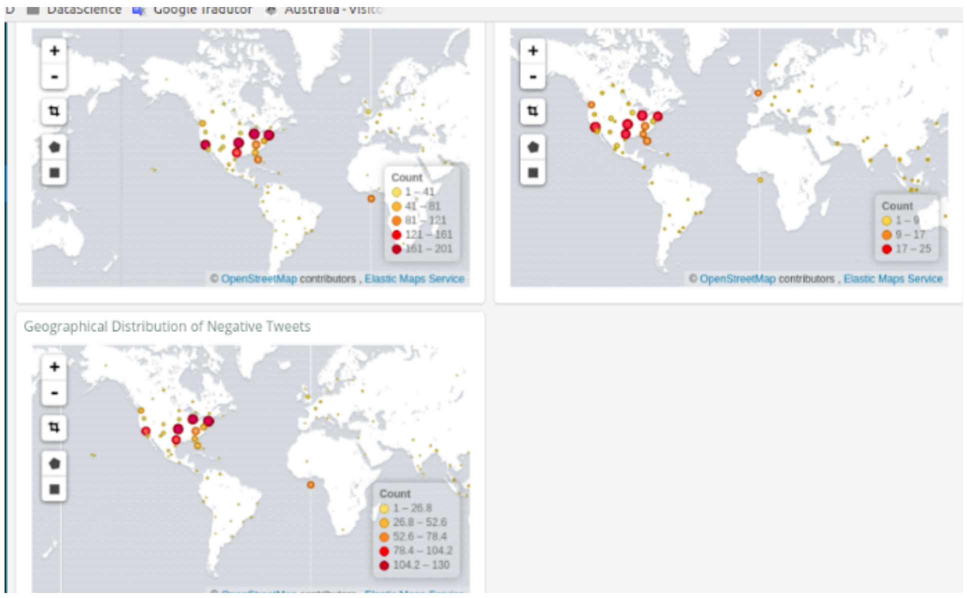
4.3. <u>Visualizer</u>: To visualize the real-time collected data you have two options:

(a) Save the data in an external json file and later show summarized results in collected data through Spark SQL.
(b) (optional, but *preferable*) Show the results in dashboards using the visualization tool Kibana. For that, you will need to:
   (i) Install ElasticSearch and Kibana.
   (ii) Create an index for visualization on elasticsearch and also in Kibana.
   (iii) Create visualization objects to show the overall analysis over tweets.
   (iv) Create a dashboard based on visualization objects.
   (v) (optional*) Create heat-maps, maps or any other geo-based graph to show the geolocation distribution of positive and negative tweets. More specifically, first heat map would show the geolocation distribution of all tweets, regardless of sentiment related to the hashtag analyzed. Second and Third heat maps would show geolocation distributions of positive tweets and negative tweets, respectively. When you send data from spark to ElasticSearch, you can add a timestamp and then in the dashboard, you set the refresh time to 2 min as an example. This way, your graphs will keep changing as you receive data in real-time.

Example of a simple Dashboard in Kibana



Example of heat-maps in a dashboard in Kibana:



Grading Criteria:

- Completing the code for cleaning the tweets: 20 points
- Applying Sentiment Analysis: 30 points
- Applying geo tool to retrieve the correct location of the tweet: 25 points
- Visualizing/Analyzing real-time collected data:
    - (Option 1) Using Spark SQL: 25 points;
    - (Option 2) Using dashboard in Kibana: 25 points (+ 20 extra points);
    - (Option 3) Using heat-maps in a dashboard in Kibana: 25 points (+ 40 extra points).

Notes:

(1) You can work in groups of at most 4 students. However, all students will need to be present in the project demonstration.

(2) In the case you choose to work with Spark SQL for the Visualization step, the grader may ask you to run any Spark SQL queries during the demonstration.

(3) Questions during the demonstration will be directed to the students and not to the group. Therefore, all team members should work hard on the tasks.

References for the Project:

For getting Geolocation of the tweet:

https://pypi.org/project/python-gmaps/ # For GoogleMaps
https://pypi.org/project/geopy/ # For Geopy

For Sentiment Analysis:

https://medium.com/@b.terryjack/nlp-pre-trained-sentiment-analysis-1eb52a9d742c

For downloading and installing Elasticsearch:

https://www.elastic.co/downloads/elasticsearch

Elasticsearch for beginners:

https://medium.com/naukri-engineering/elasticsearch-tutorial-for-beginners-using-python-b9cb48edcedc

Official Low-level client for ElasticSearch on Python:

https://pypi.org/project/elasticsearch/

For downloading and installing Kibana :

https://www.elastic.co/downloads/kibana

For creating indexes inside Kibana:

https://www.elastic.co/guide/en/kibana/current/tutorial-define-index.html

For GeoShape datatype:

https://www.elastic.co/guide/en/elasticsearch/reference/current/geo-shape.html