

**Report: CS6847: Cloud Computing Assignment 5**

**Ashish Gupta**

**CS19M012**

## I) Zookeeper Setup:

1. Create the directory in data folder which stores all the data related to zookeeper in that directory and also set the necessary ownership.

```
cs19m012@final-instance-cc:~$ sudo mkdir -p /data/zookeeper
cs19m012@final-instance-cc:~$ chown -R zookeeper:zookeeper /data/zookeeper
chown: changing ownership of '/data/zookeeper': Operation not permitted
cs19m012@final-instance-cc:~$ sudo chown -R zookeeper:zookeeper /data/zookeeper
```

2. Extract the downloaded zookeeper zip file using the following command.

```
cs19m012@final-instance-cc:~$ tar -zxvf apache-zookeeper-3.6.1-bin.tar.gz
```

3. Rename the folder to zookeeper

```
cs19m012@final-instance-cc:~$ mv apache-zookeeper-3.6.1-bin zookeeper
```

4. Create a zoo.cfg in zookeeper/conf directory and add the following configurations.

```
tickTime=2500
dataDir=/data/zookeeper
clientPort=2181
maxClientCnxns=80
```

5. Start the zookeeper service.

```
zookeeper@final-instance-cc:/home/cs19m012/zookeeper$ bin/zkServer.sh start
ZooKeeper JMX enabled by default
Using config: /home/cs19m012/zookeeper/bin/../conf/zoo.cfg
Starting zookeeper ... STARTED
```

6. Connect to local Zookeeper server using the following command:  
bin/zkCli.sh -server 127.0.0.1:2181

```

WATCHER::

WatchedEvent state:SyncConnected type:None path:null
[zk: 127.0.0.1:2181(CONNECTING) 0]
[zk: 127.0.0.1:2181(CONNECTED) 0]
[zk: 127.0.0.1:2181(CONNECTED) 0]
[zk: 127.0.0.1:2181(CONNECTED) 0] help
ZooKeeper -server host:port -client-configuration properties-file cmd args
    addWatch [-m mode] path # optional mode is one of [PERSISTENT, PERSISTENT_RECURSIVE] - default is PERSISTENT_RECURSIVE
    addauth scheme auth
    close
    config [-c] [-w] [-s]
    connect host:port
    create [-s] [-e] [-c] [-t ttl] path [data] [acl]
    delete [-v version] path
    deleteall path [-b batch size]
    delquota [-n|-b] path
    get [-s] [-w] path
    getAcl [-s] path
    getAllChildrenNumber path
    getEphemerals path
    history
    listquota path
    ls [-s] [-w] [-R] path
    printwatches on|off
    quit
    reconfig [-s] [-v version] [[-file path] | [-members serverID=host:port1:port2;port3[,...]*] | [-add serverId=host:port1:port2;port3[,...]]* [-remove serverId[,...]*]
    redo cmdno
    removewatches path [-c|-d|-a] [-l]
    set [-s] [-v version] path data
    setAcl [-s] [-v version] [-R] path acl
    setquota -n|-b val path
    stat [-w] path
    sync path
    version
Command not found: Command not found help

```

7. Create a systemd service for zookeeper by creating a file in /etc/systemd/system

```

[Unit]
Description=Zookeeper Daemon
Documentation=http://zookeeper.apache.org
Requires=network.target
After=network.target

[Service]
Type=forking
WorkingDirectory=/home/cs19m012/zookeeper
User=zookeeper
Group=zookeeper
ExecStart=/home/cs19m012/zookeeper/bin/zkServer.sh start /home/cs19m012/zookeeper/conf/zoo.cfg
ExecStop=/home/cs19m012/zookeeper/bin/zkServer.sh stop /home/cs19m012/zookeeper/conf/zoo.cfg
ExecReload=/home/cs19m012/zookeeper/bin/zkServer.sh restart /home/cs19m012/zookeeper/conf/zoo.cfg
TimeoutSec=30
Restart=on-failure

[Install]
WantedBy=default.target

```

8. Reload the daemon using the following command:

Systemctl daemon reload

## II) Kafka Setup:

1. Copy the downloaded tgz file to the VM and unzip the tgz file.

```
cs19m012@final-instance-cc:~$ tar zxvf kafka_2.12-2.5.0.tgz
kafka_2.12-2.5.0/
kafka_2.12-2.5.0/LICENSE
kafka_2.12-2.5.0/NOTICE
kafka_2.12-2.5.0/bin/
kafka_2.12-2.5.0/bin/kafka-delete-records.sh
kafka_2.12-2.5.0/bin/trogdor.sh
kafka_2.12-2.5.0/bin/kafka-preferred-replica-election.sh
kafka_2.12-2.5.0/bin/connect-mirror-maker.sh
kafka_2.12-2.5.0/bin/kafka-console-consumer.sh
kafka_2.12-2.5.0/bin/kafka-consumer-perf-test.sh
kafka_2.12-2.5.0/bin/kafka-log-dirs.sh
kafka_2.12-2.5.0/bin/zookeeper-server-stop.sh
kafka_2.12-2.5.0/bin/kafka-verifiable-consumer.sh
kafka_2.12-2.5.0/bin/kafka-acls.sh
kafka_2.12-2.5.0/bin/zookeeper-server-start.sh
kafka_2.12-2.5.0/bin/kafka-server-stop.sh
kafka_2.12-2.5.0/bin/kafka-configs.sh
kafka_2.12-2.5.0/bin/kafka-reassign-partitions.sh
kafka_2.12-2.5.0/bin/kafka-leader-election.sh
kafka_2.12-2.5.0/bin/kafka-producer-perf-test.sh
kafka_2.12-2.5.0/bin/kafka-topics.sh
kafka_2.12-2.5.0/bin/connect-standalone.sh
kafka_2.12-2.5.0/bin/kafka-dump-log.sh
kafka_2.12-2.5.0/bin/kafka-broker-api-versions.sh
kafka_2.12-2.5.0/bin/kafka-consumer-groups.sh
kafka_2.12-2.5.0/bin/connect-distributed.sh
kafka_2.12-2.5.0/bin/kafka-delegation-tokens.sh
kafka_2.12-2.5.0/bin/kafka-run-class.sh
kafka_2.12-2.5.0/bin/kafka-replica-verification.sh
kafka_2.12-2.5.0/bin/kafka-console-producer.sh
kafka_2.12-2.5.0/bin/zookeeper-shell.sh
kafka_2.12-2.5.0/bin/windows/
kafka_2.12-2.5.0/bin/windows/kafka-log-dirs.bat
kafka_2.12-2.5.0/bin/windows/zookeeper-server-stop.bat
```

2. Modify the zookeeper.service file.

```
cs19m012@final-instance-cc:~$ cat /etc/systemd/system/zookeeper.service
[Unit]
Description=Zookeeper Daemon
Documentation=http://zookeeper.apache.org
Requires=network.target remote-fs.target
After=network.target remote-fs.target

[Service]
#Type=forking
Type=simple
Environment="JAVA_HOME=/home/cs19m012/jdk1.8.0_241"
#WorkingDirectory=/home/cs19m012/zookeeper
User=cs19m012
Group=cs19m012
#ExecStart=/home/cs19m012/zookeeper/bin/zkServer.sh start /home/cs19m012/zookeeper/conf/zoo.cfg
#ExecStop=/home/cs19m012/zookeeper/bin/zkServer.sh stop /home/cs19m012/zookeeper/conf/zoo.cfg
#ExecReload=/home/cs19m012/zookeeper/bin/zkServer.sh restart /home/cs19m012/zookeeper/conf/zoo.cfg
ExecStart=/home/cs19m012/kafka/bin/zookeeper-server-start.sh /home/cs19m012/kafka/config/zookeeper.properties
ExecStop=/home/cs19m012/kafka/bin/zookeeper-server-stop.sh
Restart=on-abnormal
#TimeoutSec=30
#Restart=on-failure

[Install]
#WantedBy=default.target
WantedBy=multi-user.targetcs19m012@final-instance-cc:~$
```

### 3. Create a kafka.service file

```
cs19m012@final-instance-cc:~$ cat /etc/systemd/system/kafka.service
[Unit]
Description=Apache Kafka Server
Documentation=http://kafka.apache.org/documentation.html
Requires=zookeeper.service
After=zookeeper.service

[Service]
Type=simple
Environment="JAVA_HOME=/home/cs19m012/jdk1.8.0_241"
ExecStart=/home/cs19m012/kafka/bin/kafka-server-start.sh /home/cs19m012/kafka/config/server.properties
ExecStop=/home/cs19m012/kafka/bin/kafka-server-stop.sh

[Install]
WantedBy=multi-user.target
```

### 4. Reload the daemon using the following command.

```
sudo systemctl daemon-reload
```

### 5. Start the zookeeper service first always and then start the kafka service.

```
sudo systemctl start zookeeper
```

```
sudo systemctl start kafka
```

### 6. The topic can be created using the following command:

```
bin/kafka-topics.sh --create --zookeeper localhost:2181 --
replication-factor 1 --partitions 1 --topic testTopic
```

## Creating web application using Python

### Prerequisite Installation:

#### 1. Install flask using command: pip install flask

```

cs19m012@final-instance-cc:~$ pip install flask
Collecting flask
  Downloading https://files.pythonhosted.org/packages/f2/28/2a03252dfb9ebf377f40fba6a7841b47083260bf8bd8e737b0c6952df83f/Flask-1.1.2-py2.py3-none-any.whl (94kB)
    100% |#####| 102kB 2.9MB/s
Collecting click>=5.1 (from flask)
  Downloading https://files.pythonhosted.org/packages/d2/3d/fa76db83bf75c4f8d338c2fd15c8d33fdd7ad23a9b5e57eb6c5de26b430e/click-7.1.2-py2.py3-none-any.whl (82kB)
    100% |#####| 92kB 6.1MB/s
Collecting Werkzeug>=0.15 (from flask)
  Downloading https://files.pythonhosted.org/packages/cc/94/5f7079a0e00bd6863ef8flda638721e9da21e5bacee597595b318f71d62e/Werkzeug-1.0.1-py2.py3-none-any.whl (298kB)
    100% |#####| 307kB 2.8MB/s
Collecting itsdangerous>=0.24 (from flask)
  Downloading https://files.pythonhosted.org/packages/76/ae/44b03b253d6fade317f32c24dl00b3b35c2239807046a4c953c7b89fa49e/itsdangerous-1.1.0-py2.py3-none-any.whl
Collecting Jinja2>=2.10.1 (from flask)
  Downloading https://files.pythonhosted.org/packages/30/9e/f663a2aa66a09d838042aela2c5659828bb9b41ea3a6efa20a20fd92b121/Jinja2-2.11.2-py2.py3-none-any.whl (125kB)
    100% |#####| 133kB 4.8MB/s

```

### III) Topic Creation in Kafka:

Create topics named NY and CA using the following commands:

```
bin/kafka-topics.sh --create --zookeeper localhost:2181 --
replication-factor 1 --partitions 1 --topic NY
```

```

cs19m012@final-instance-cc:~/kafka$ bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic NY
Created topic NY.
cs19m012@final-instance-cc:~/kafka$ bin/kafka-topics.sh --create --zookeeper localhost:2181 --replication-factor 1 --partitions 1 --topic CA
Created topic CA.

```

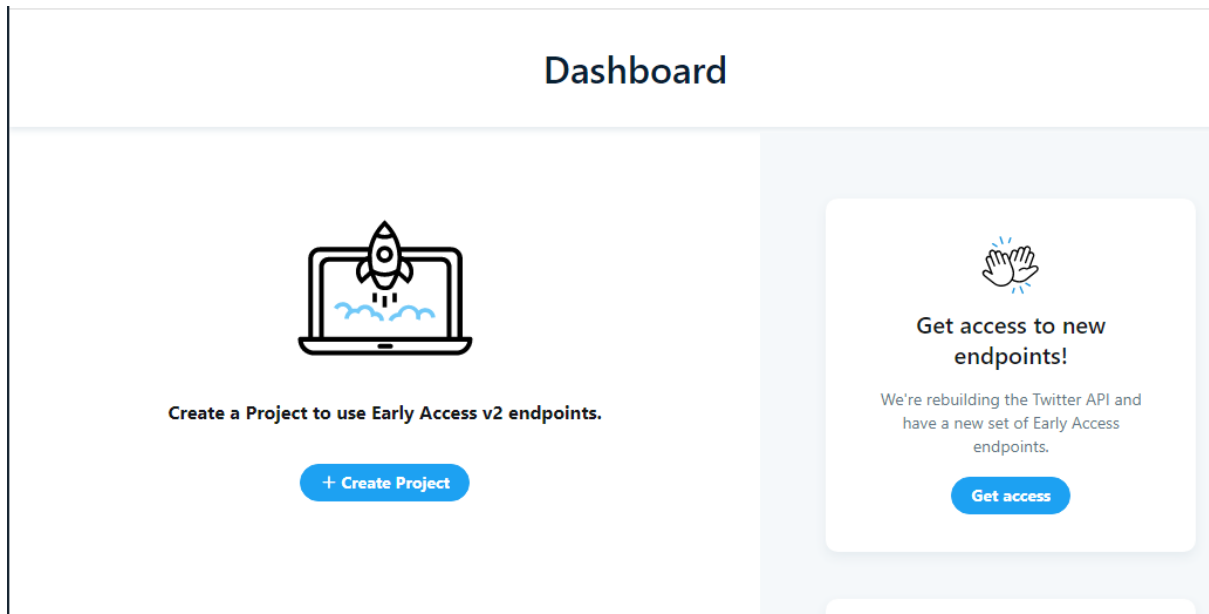
### IV) Spark Stream Setup:

1. Spark Streaming API enables scalable, high-throughput, fault-tolerant stream processing of live data streams.
2. To integrate Spark streaming API with Kafka, we need libraries of Spark streaming that should be added in the classpath while we try to run our streaming program. The library that we use are “**spark-streaming-kafka-0-8-assembly\_2.11-2.4.5.jar**” file.
3. This JAR file can be downloaded from Maven Central <http://search.maven.org/>. Specific details of the version is given below:  
**Group Id = org.apache.spark,**  
**Artifact Id = spark-streaming-kafka-0-8-assembly,**  
**Version = 2.4.5.**
4. Once downloaded, we can use the JAR file in the spark submit command as follows:


```
$SPARK_HOME/bin/spark-submit --jars spark-streaming-kafka-0-8-assembly_2.11-2.4.5.jar --packages org.apache.spark:spark-streaming-kafka-0-10_2.12:2.4.5 twitter_stream.py
```

## V) Twitter API Setup:

1. Apply for Twitter Developer Account. Once request has been approved, you can create a new project by clicking on Create Project button on Dashboard.



2. Name your Project.



## Name your Project

Your Project helps you organize your work and monitor your usage with the Twitter API.

STEP 1

63

Next

3. Select a use case.



## Which best describes you?

Below you will find options on how you intend to use the Twitter Developer Platform.

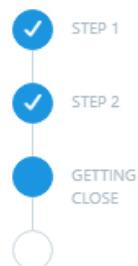
Student



Next

Back

4. Describe your project.



## Describe your new Project

This info is just for us, here at Twitter. It'll help us create better developer experiences down the road.

End semester assignment for Cloud computing



Next

Back



5. Name your App.



- ✓ STEP 1
- ✓ STEP 2
- ✓ GETTING CLOSE
- ALMOST DONE!

## Last step, name your App

Apps are where you get your **access keys and tokens** and set permissions. They are encompassed within your Projects.

16

Complete

Back

6. Copy and save the API key and API secret key in a text file.

## Here are your keys & tokens

For security, this will be the last time we'll display these. If something happens, you can always regenerate them.

### API key ⓘ

iIK07UyN19L9hgFKMZOFaaFDx



### API secret key ⓘ

lTr4KzPxLBNiqwCodXh8vuJ1hf0twnehH9ZCasfMOfis67rsrV



### Bearer token ⓘ

AAAAAAAAAAAAAAAAAAAAAAAAANwSOAEAAAAAZqvHQswlsKx2xZZfeaHeGC3  
sEDw%3D7yTyup1adaK1ODnebwZf6CmdacLeBKrOAG2RL7Ov2fm2hjcwYF



7. After clicking on Keys and tokens tab under Settings on Project, generate Access token and secret by clicking on **Generate** near it. Save them in a same text file where API key and secret key was saved.

Here are your new access token & secret. Have you saved them?



For security, this will be the last time we'll display these. If something happens, you can always regenerate them.

Access token: 1257707359213780992-m5nQDJfZf9zWcuYD8NjoentkSXd2Ym



Access token secret: xIFafx7iY2lj8vRC1SOPhLBDJsQFKccbj8PIKqFvj5hgG



Yes, I saved them

## VI) Publishing tweets to Kafka:

1. Since the topics NY and CA are already created previously in Kafka, we can invoke the script that will be fetch the tweets based on the location NY and CA and store them in their respective topics.
2. Initially, it creates a stream to read the tweets and then publish them to Kafka using KafkaClient and SimpleProducer.
3. To do so, we first invoke the command: **python read\_tweets.py NY**

```
ps19m012@final-instance-cc:~$ python read_tweets.py NY
Today is my birthday, y'all! 🎂 🍰 🍰 I'm 28 now, and it feels bitter sweet. Existential angst is creeping up on m
e, b_ https://t.co/uVBCJunFWI
W Stream !! Luv u niggas 🍷 🍷
Tbh I would've preferred if weed just got completely decriminalized instead of legalized
First one in the books #Warzone #CallOfDuty https://t.co/Dg0Vclpzhe
Perfect right now.
Of course the *real* way to allude to his role would be if Jonah cut his nipple off and presented it to Amy as proo
https://t.co/a9oggj8rcJ
@Kitacookie plz i'm having a bad few days
Oat Milk Latte at Ralph's and some @Chiefs and @DCSuperman gear 😊

#zacksnydersjusticeleague #Superman_ https://t.co/cdVQWCPu98
The most perfect spring day to sit in the park and listen to jazz. I ❤️ NY https://t.co/DSkzf6KkzT
I've been growing my hair for so long the bottom is all sun bleached while the new growth is almost black looking l
https://t.co/fUV90TmPj2
What are the Jets going to do at 2?

Who's the QB going to be Wilson or Fields?

Will the 49ers get their man?
interesting... loop to the topic we discussed @noname_8
@JennRavenna I literally pray every day to have the confidence that white people have to just drop flaming caca tak
https://t.co/csaCugglLe
New York Is Likely to Legalize Marijuana. Here's What to Know. https://t.co/d2v6NYHrSg
Just posted a photo @ Church-in-the-Garden https://t.co/gsionLR2Uf
Updated: Construction on #Systemwide https://t.co/64GL4ojhcQ
```

## VII) Consume tweets through Kafka:

1. To check whether the tweets have reached the Kafka topic, we open a separate terminal and run the following command:

```
bin/kafka-console-consumer.sh --bootstrap-server localhost:9092 --topic NY
--from-beginning
```

```
I jus added a favorite to my do not disturb list...I'm acting unusual  
Why did I take that nap earlier?! Tf  
Good morning  
Playlist::@ciara = All Good #MasterpieceTheater 🍌👉🔵🔴📺👉🍌👉🍌👉🔵🔴📺👉🍌👉🍌👉🔵🔴📺👉🍌👉🍌👉🔵🔴📺👉 https://  
o/bws3APCQZl  
when did the new yorker become buzzfeed 😬  
@MSNBC Words have consequences #Dominion  
Y'all think Cordae be waiting with orange slices?  
#Flashback Spring/Summer 2014. Team Scan was one of the premier teams in the @nikeeyb &mp; was on the rise at that  
tim_ https://t.co/6dwJT70zvc  
i love the first ammendment and all but it should be illegal to use the word "woke" more than once in a tweet.  
If you workin on ya MD and like when my chains hit ya face tap in one time i got questions ☹️  
Don't bring your "SNL hasn't been funny since Belushi left" nonsense up in here. This is gold.  
What in the world  
@ernestdove Luchessi could be the second best starter on the staff. Consistency and health. Have him start every 5t  
h day  
Oh sure. NOW he grows a conscience!! #SSSS_GRIDMAN #Toonami  
@carl_weezer everyone agrees with you though  
@AndiHazzey @RedState But do you?  
I'm curious if.... if he pays them well or did he exploit them 😞  
SouthernPLAYALISTIC ...FEEL THE SPEAKERS IN MY ♥💙 ... 🔥🔥🔥🔥🔥🔥🔥  
Going on a day trip to Niagara Falls today and I'm so excited. All the years I've been back &mp; forth to NY, I'v  
e never been  
I BEEN DELIVERED https://t.co/HnNAcoVPS4  
Truly absurd  
This is all because "the EFFETE ELITE LEFT" (who think they run things, live in gated, guarded homes)  
Nothing else_ https://t.co/XtNKGZKuJB  
@johncardillo @CNN You buffoons obviously didn't read the article or don't know that the word "accident" is a commo  
_ https://t.co/PiOy26wUJu  
@CiemaMaMa21 @phillipmasina02 @KhulekaniKubayi @mafisto27 @PortiaShongwe2 @AbelJnr_Mhlongo @Real_sikhulu @PeEunice_  
https://t.co/hghUGU7JEx  
I'm sorry I just don't want banana milk this is weird  
@LilBoFLO Oh fuck we got a man down here  
Checkmate  
Help the Anwnars Find Peace https://t.co/QtWrsOBvBY  
@yeezuslover I'm aiming for something simple like something people could just throw a calm simple fit with if that  
makes sense  
Y'all think Cordae and Serena man be sharing orange slices? https://t.co/WhtbQtOmT5  
Just IN-at JFK loc.- @CrownePlaza @crowneplazanyc George and his sidekick Mgr, treat paying cust. like me &mp; ano  
ther_ https://t.co/twB2dbEBBS  
@karabo_27Dec Are you mad.. https://t.co/CSOlCHgtlx  
@naomiosaka I know he can't slice oranges better than I do https://t.co/70bgwk1RUQ
```

**VIII) 10 mins past tweets and count:**

1. Now, we run the tweet analysis program which will consume the tweets from Kafka, do the processing and then return the tweets as well as the count.
2. The following command should be executed for the same:

```
$SPARK_HOME/bin/spark-submit --jars spark-streaming-kafka-0-8-assembly_2.11-2.4.5.jar --packages org.apache.spark:spark-streaming-kafka-0-10_2.12:2.4.5 twitter stream.py NY
```

```

-----
Time: 2021-03-26 17:27:00
-----
Tweets in this batch: 11
-----
Time: 2021-03-26 17:27:00
-----
Total tweets till now: 11
-----
Time: 2021-03-26 17:27:30
-----
Idk why costar didnt mention someone busting into my apartment and punching me in the face today.
Thank you @TIDAL for a a running playlist today that I didnt know I needed.
"stop crying it's just a movie."

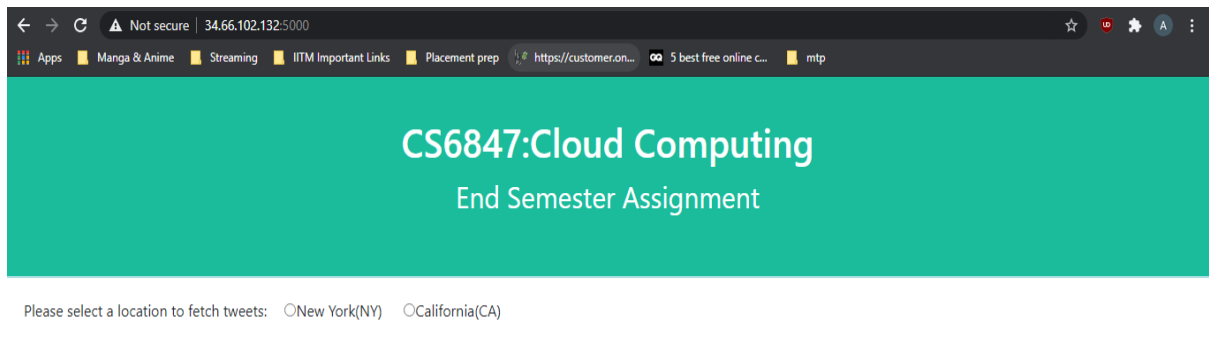
the movie: https://t.co/uRWimLOS1w
@marisairi_LMFAOO
@Angelina161813 @nbcsnl I dont think even SHE can save the show
@DannyGroner For me, the Wikipedia information is important & connotes credibility.
Yall remember those sleep over lock-ins @ church ?
Easter Phoebe at The Pool just now. Lifer for me! #birdcpp #lifer @NYCAudubon @My_Cen_ParkNYC @CentralPark_NYC http
s://t.co/tiuiG2998s
Its crack the window open weather outside
Just posted a video @ Warren Park https://t.co/CcoIz1zcDZ
...
-----
Time: 2021-03-26 17:27:30
-----
Tweets in this batch: 11
-----
Time: 2021-03-26 17:27:30
-----
Total tweets till now: 22

```

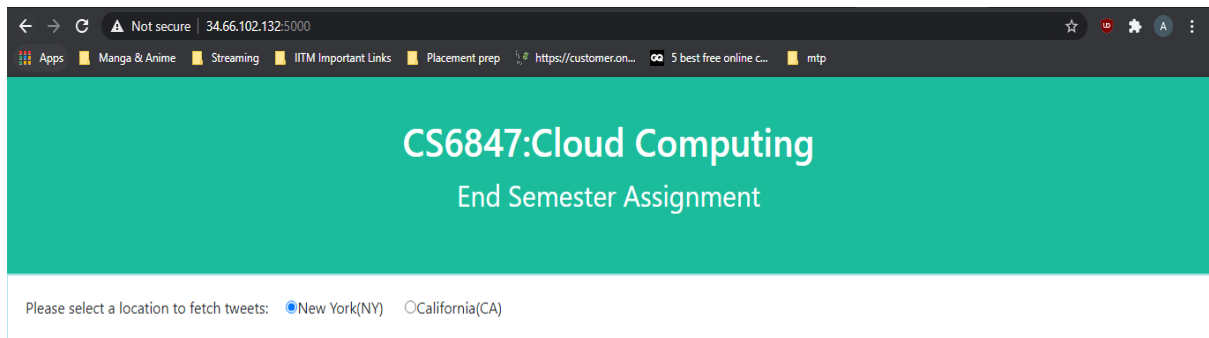
## IX) Web Interface:

1. I have used Flask to build web interface in Python that is deployed over Cloud and can be accessed on port number 5000 after typing the URL:

<http://<VM Instance Public IP>:5000/>



Total tweets till now:



Total tweets till now: 180

#### X) **Conclusion:**

1. In this final assignment, I was able to install and configure Kafka and Zookeeper.
2. Since the assignment had several modules, I was able to understand the basics of each module and was able to integrate Kafka with Spark and also use Twitter API to fetch tweets and do the processing using Spark.
3. Also, I was able to implement a Web interface using Flask in Python to display the results of processing.