



PRINCIPLES OF BIGDATA MANAGEMENT
PHASE 1

TEAM MEMBERS

AVINASH GANGURI 16293133

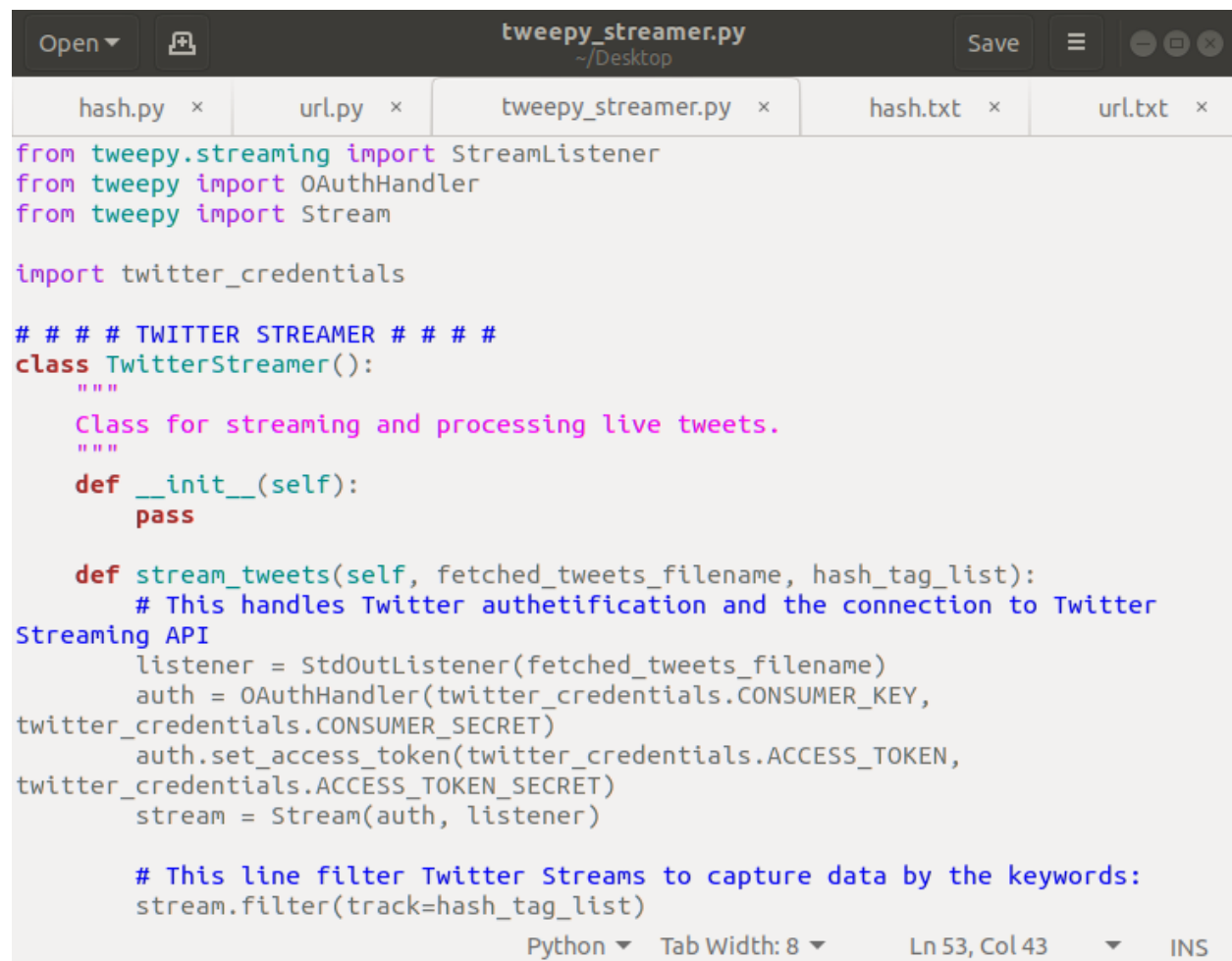
SRI SAI NIKHIL KANTIPUDI 16292907

ROSHNA TOKE 16293711

The main objective of this project is to collect twitter data by using twitter API and the extracting hashtags and URLs from the data and running wordcount on Hadoop and spark.

1. Collecting twitter data using twitter API
2. Extracting hashtags and URLs
3. Running wordcount program in Hadoop and spark

Collecting twitter data using twitter API:



```
from tweepy.streaming import StreamListener
from tweepy import OAuthHandler
from tweepy import Stream

import twitter_credentials

##### TWITTER STREAMER #####
class TwitterStreamer():
    """
    Class for streaming and processing live tweets.
    """
    def __init__(self):
        pass

    def stream_tweets(self, fetched_tweets_filename, hash_tag_list):
        # This handles Twitter authentication and the connection to Twitter
        # Streaming API
        listener = StdOutListener(fetched_tweets_filename)
        auth = OAuthHandler(twitter_credentials.CONSUMER_KEY,
                             twitter_credentials.CONSUMER_SECRET)
        auth.set_access_token(twitter_credentials.ACCESS_TOKEN,
                             twitter_credentials.ACCESS_TOKEN_SECRET)
        stream = Stream(auth, listener)

        # This line filter Twitter Streams to capture data by the keywords:
        stream.filter(track=hash_tag_list)
```

Twitter data extraction:

```
sri@SRI: ~/hadoop
File Edit View Search Terminal Help
6279990969978880"},"indices":[0,9]]},"symbols":[],"favorited":false,"retweeted":false,"filter_level":"low","lang":"fr","timestamp_ms":"1568568303384"}

{"created_at":"Sun Sep 15 17:25:03 +0000 2019","id":1173286619648012288,"id_str":"1173286619648012288","text":"THIS JOURNALIST IS HIGHLIGHTING TWITCH STUFF!!! Talk to her","source":"\u003ca href=\"http://twitter.com/download/iphone\" rel=\"nofollow\"\u003eTwitter for iPhone\u003c/a\u003e","truncated":false,"in_reply_to_status_id":null,"in_reply_to_status_id_str":null,"in_reply_to_user_id":null,"in_reply_to_user_id_str":null,"in_reply_to_screen_name":null,"user":{"id":1072587960141406211,"id_str":"1072587960141406211","name":"ralphzzz","screen_name":"ralphzzzz13","location":null,"url":null,"description":"used to have favourites....","translator_type":"none","protected":false,"verified":false,"followers_count":61,"friends_count":663,"listed_count":0,"favourites_count":17166,"statuses_count":3052,"created_at":"Tue Dec 11 20:24:32 +0000 2018","utc_offset":null,"time_zone":null,"geo_enabled":true,"lang":null,"contributors_enabled":false,"is_translator":false,"profile_background_color":"F5F8FA","profile_background_image_url":"","profile_background_image_url_https":"","profile_background_tile":false,"profile_link_color":"1DA1F2","profile_sidebar_border_color":"C0DEED","profile_sidebar_fill_color":"DDEEF6","profile_text_color":"333333","profile_use_background_image":true,"profile_image_url":"http://pbs.twimg.com/profile_images/1140771293094170624/WvcS7NZ5_normal.jpg","profile_image_url_https":"http://pbs.twimg.com/profile_images/1140771293094170624/WvcS7NZ5_normal.jpg","default_profile":true,"default_profile_image":false,"following":null,"follow_request_sent":null,"notifications":null},"geo":null,"coordinates":null,"place":null,"contributors":null,"quoted_status_id":1173273726445658113,"quoted_status_id_str":"1173273726445658113","quoted_status":{"created_at":"Sun Sep 15 16:33:49 +0000 2019","id":1173273726445658113,"id_str":"1173273726445658113","text":"Friends! I'm going to be at @TwitchCon at the end of the month! If you're a streamer/gamer/influencer/internet hum\u0026 https://t.co/vE8b1B7KmlI \"source
```

Keyword used for collecting twitter data


- Cricket
- ISRO
- a
- b

Extracting hashtags and URLs from collected twitter data:

As the twitter data is in the text file. We need to send the text file as the input for the python code which extracts the hashtags and urls.

```
$>python3 hash.py >> hash.txt
```

```
$>python3 hash.py >> url.txt
```



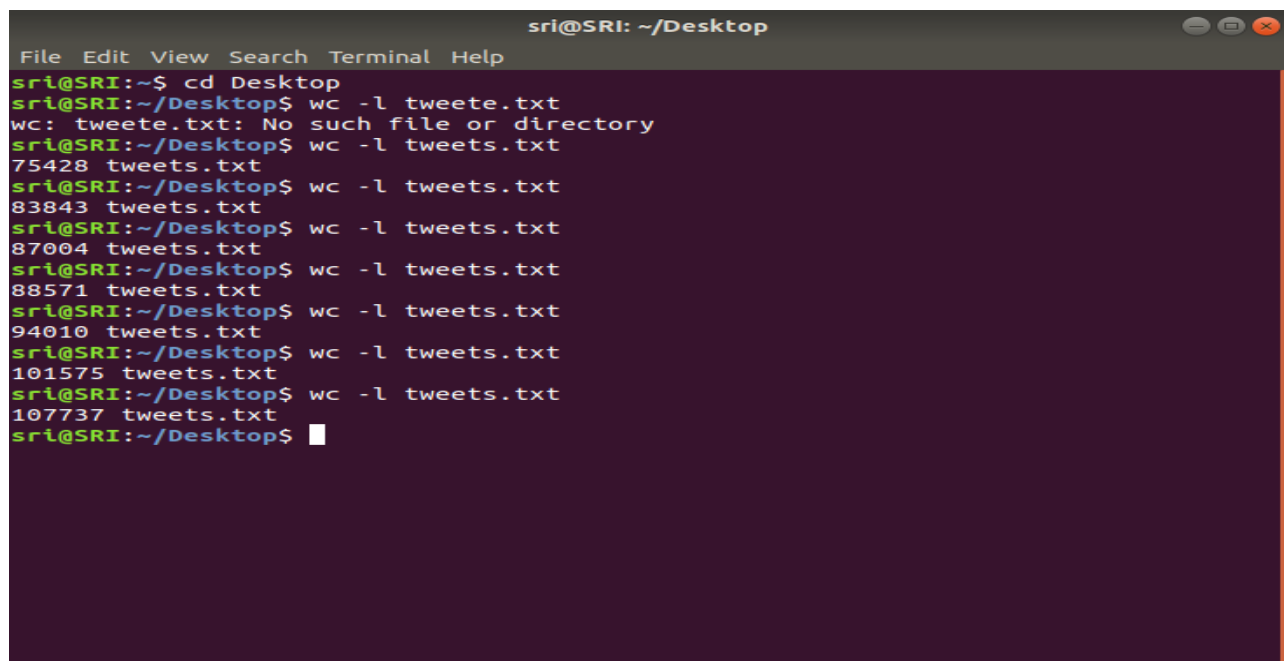
```
import json

data = list(open('tweet|.txt', 'rt'))

for x in data:
    x = json.loads(x)
    try:
        if len(x['entities']['hashtags']) != 0:
            for p in x['entities']['hashtags']:
                print(p['text'])
    except KeyError:
        pass
```

hash.txt and url.txt are the text files that contains hashtags and urls. We need to run the wordcount program to count the words.

Total tweets collected: 107737



```
sri@SRI: ~/Desktop
File Edit View Search Terminal Help
sri@SRI:~$ cd Desktop
sri@SRI:~/Desktop$ wc -l tweete.txt
wc: tweete.txt: No such file or directory
sri@SRI:~/Desktop$ wc -l tweets.txt
75428 tweets.txt
sri@SRI:~/Desktop$ wc -l tweets.txt
83843 tweets.txt
sri@SRI:~/Desktop$ wc -l tweets.txt
87004 tweets.txt
sri@SRI:~/Desktop$ wc -l tweets.txt
88571 tweets.txt
sri@SRI:~/Desktop$ wc -l tweets.txt
94010 tweets.txt
sri@SRI:~/Desktop$ wc -l tweets.txt
101575 tweets.txt
sri@SRI:~/Desktop$ wc -l tweets.txt
107737 tweets.txt
sri@SRI:~/Desktop$
```

Wordcount on Hadoop:

Make sure that the data node and name node are running. Otherwise start them

```
$> ./start-dfs.sh
```

```
$> ./start-yarn.sh
```

For hashtags:

```
$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar  
wordcount /nittest/hash.txt /nittest/out
```

```
$ hdfs dfs -ls /nittest/out/
```

```
$ hdfs dfs -cat /nittest/out/part-r-00000
```

```
sri@SRI: ~/hadoop
File Edit View Search Terminal Help
-rw-r--r--  1 sri supergroup    117735 2019-09-15 13:06 /nittest/out/part-r-00000
sri@SRI:~/hadoop$ ^C
sri@SRI:~/hadoop$ hdfs dfs -cat /nittest/out/part-r-00000

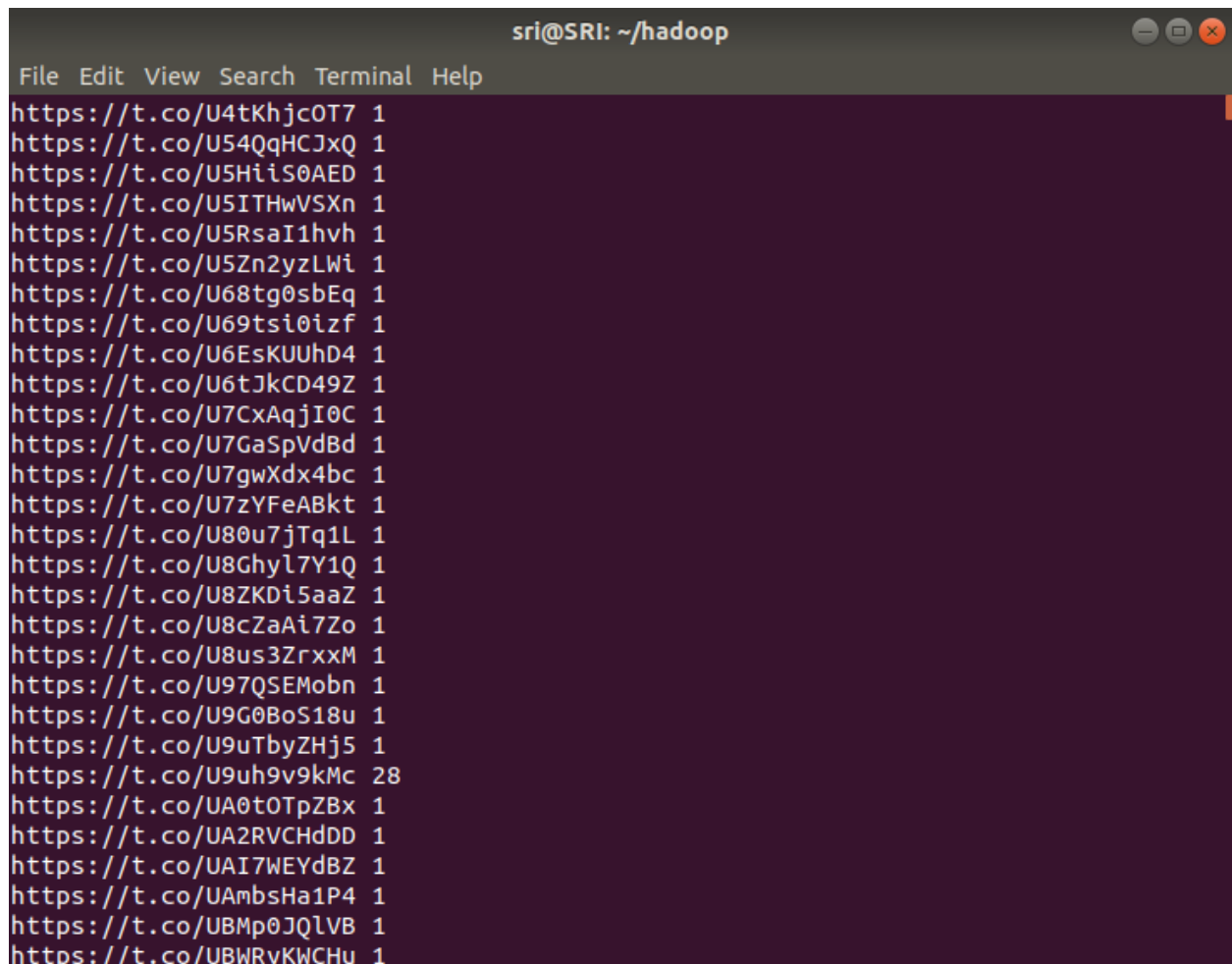
19/09/15 13:08:45 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable
100DaysNoVikas      2
100DaysOfCode       2
100MostBeautifulFaces2019      5
100MostHandsomeFaces2019      4
100daysofgratitude      1
100faces             1
100mostbeautifulfaces2019      2
101AñosCONCAMIN      1
10K                  2
10Medidas            1
10RTされたら3日間性転換RTした人見た人もやる      1
10ThingsAboutChimamanda      1
12Sep               1
12gang              1
13Sep               2
13sep2019           1
14Sep              11
14Sept              3
14septiembre        1
15DeSeptiembre      17
15Eylül1918         1
15S                 2
```

For URLs:

```
$ hadoop jar share/hadoop/mapreduce/hadoop-mapreduce-examples-2.9.0.jar  
wordcount /nittest/url.txt /nittest/outurl
```

```
$ hdfs dfs -ls /nittest/outurl/
```

```
$ hdfs dfs -cat /nittest/outurl/part-r-00000
```



```
sri@SRI: ~/hadoop
File Edit View Search Terminal Help
https://t.co/U4tKhjcOT7 1
https://t.co/U54QqHCJxQ 1
https://t.co/U5HiiS0AED 1
https://t.co/U5ITHwVSXn 1
https://t.co/U5RsaI1hvh 1
https://t.co/U5Zn2yzLWi 1
https://t.co/U68tg0sbEq 1
https://t.co/U69tsi0izf 1
https://t.co/U6EsKUUhD4 1
https://t.co/U6tJkCD49Z 1
https://t.co/U7CxAqjI0C 1
https://t.co/U7GaSpVdBd 1
https://t.co/U7gwXdx4bc 1
https://t.co/U7zYFeABkt 1
https://t.co/U80u7jTq1L 1
https://t.co/U8Ghyl7Y1Q 1
https://t.co/U8ZKD15aaZ 1
https://t.co/U8cZaAi7Zo 1
https://t.co/U8us3ZrxxM 1
https://t.co/U97QSEMobn 1
https://t.co/U9G0BoS18u 1
https://t.co/U9uTbyZHj5 1
https://t.co/U9uh9v9kMc 28
https://t.co/UA0tOTpZBx 1
https://t.co/UA2RVChdDD 1
https://t.co/UAI7WEYdBZ 1
https://t.co/UAmbSHa1P4 1
https://t.co/UBMp0JQLVB 1
https://t.co/UBWRvKWCHu 1
```

Wordcount using spark:

For hashtags:

```
$SPARK_HOME/bin/spark-submit run-example JavaWordCount /nittest/hash.txt
```

```
sri@SRI: ~  
File Edit View Search Terminal Help  
sri@SRI:~$ $SPARK_HOME/bin/spark-submit run-example JavaWordCount /nittest/hash.  
txt  
19/09/15 14:34:05 INFO spark.SparkContext: Running Spark version 2.2.0  
19/09/15 14:34:07 WARN util.NativeCodeLoader: Unable to load native-hadoop libr  
ary for your platform... using builtin-java classes where applicable  
19/09/15 14:34:08 WARN util.Utils: Your hostname, SRI resolves to a loopback ad  
dress: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)  
19/09/15 14:34:08 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to an  
other address  
19/09/15 14:34:08 INFO spark.SparkContext: Submitted application: JavaWordCount  
19/09/15 14:34:08 INFO spark.SecurityManager: Changing view acls to: sri  
19/09/15 14:34:08 INFO spark.SecurityManager: Changing modify acls to: sri  
19/09/15 14:34:08 INFO spark.SecurityManager: Changing view acls groups to:  
19/09/15 14:34:08 INFO spark.SecurityManager: Changing modify acls groups to:  
19/09/15 14:34:08 INFO spark.SecurityManager: SecurityManager: authentication d  
isabled; ui acls disabled; users with view permissions: Set(sri); groups with  
view permissions: Set(); users with modify permissions: Set(sri); groups with  
modify permissions: Set()  
19/09/15 14:34:10 INFO util.Utils: Successfully started service 'sparkDriver' o  
n port 41643.  
19/09/15 14:34:10 INFO spark.SparkEnv: Registering MapOutputTracker  
19/09/15 14:34:10 INFO spark.SparkEnv: Registering BlockManagerMaster  
19/09/15 14:34:10 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spa  
rk.storage.DefaultTopologyMapper for getting topology information  
19/09/15 14:34:10 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEn  
dpoint up  
19/09/15 14:34:10 INFO storage.DiskBlockManager: Created local directory at /tm  
p/blockmgr-c95e1e74-010a-4ea1-90ce-2b90b57d7d4c  
19/09/15 14:34:10 INFO memory.MemoryStore: MemoryStore started with capacity 36
```

```
sri@SRI: ~  
File Edit View Search Terminal Help  
WOLCHE: 1  
JessicaNkosi: 1  
moda: 2  
ADOPT: 1  
BatmanDay: 1  
SeniorCare: 1  
CassiniHuygens: 1  
httr: 1  
Ch_NCT: 1  
US: 4  
Carmel: 1  
15sep: 2  
LibDems: 8  
Ardiente: 1  
hashtags: 1  
EmillyAraújo: 1  
AapleSarkar: 1  
EasternNavalCommand: 1  
NahitDuru: 1  
nfljapan: 1  
Kagame: 1  
NYCFC: 1  
ondragonwingsstudio: 1  
DemonSlayer: 1  
CountryMusicPBS: 3  
RenewableEnergy: 1  
MehdiHasan: 1  
Budapest: 1  
WeAreNewYorksTeam: 1
```


For URLs:

`$SPARK_HOME/bin/spark-submit run-example JavaWordCount /nittest/url.txt`

```
sri@SRI: ~  
File Edit View Search Terminal Help  
sri@SRI:~$  
sri@SRI:~$  
sri@SRI:~$  
sri@SRI:~$ $SPARK_HOME/bin/spark-submit run-example JavaWordCount /nittest/url.txt  
19/09/15 14:32:38 INFO spark.SparkContext: Running Spark version 2.2.0  
19/09/15 14:32:40 WARN util.NativeCodeLoader: Unable to load native-hadoop library for your platform... using builtin-java classes where applicable  
19/09/15 14:32:41 WARN util.Utils: Your hostname, SRI resolves to a loopback address: 127.0.1.1; using 10.0.2.15 instead (on interface enp0s3)  
19/09/15 14:32:41 WARN util.Utils: Set SPARK_LOCAL_IP if you need to bind to another address  
19/09/15 14:32:41 INFO spark.SparkContext: Submitted application: JavaWordCount  
19/09/15 14:32:41 INFO spark.SecurityManager: Changing view acls to: sri  
19/09/15 14:32:41 INFO spark.SecurityManager: Changing modify acls to: sri  
19/09/15 14:32:41 INFO spark.SecurityManager: Changing view acls groups to:  
19/09/15 14:32:41 INFO spark.SecurityManager: Changing modify acls groups to:  
19/09/15 14:32:41 INFO spark.SecurityManager: SecurityManager: authentication disabled; ui acls disabled; users with view permissions: Set(sri); groups with view permissions: Set(); users with modify permissions: Set(sri); groups with modify permissions: Set()  
19/09/15 14:32:43 INFO util.Utils: Successfully started service 'sparkDriver' on port 38583.  
19/09/15 14:32:43 INFO spark.SparkEnv: Registering MapOutputTracker  
19/09/15 14:32:44 INFO spark.SparkEnv: Registering BlockManagerMaster  
19/09/15 14:32:44 INFO storage.BlockManagerMasterEndpoint: Using org.apache.spark.storage.DefaultTopologyMapper for getting topology information  
19/09/15 14:32:44 INFO storage.BlockManagerMasterEndpoint: BlockManagerMasterEndpoint up
```

```
sri@SRI: ~  
File Edit View Search Terminal Help  
https://t.co/XsmsmTKYMJ: 1  
https://t.co/mY59dtlNMQ: 1  
https://t.co/3c2iD6oCF3: 1  
https://t.co/LBN5pDFs3q: 1  
https://t.co/DlKC1zS1qT: 1  
https://t.co/1EjtQM294k: 1  
https://t.co/7Lma32i9E7: 1  
https://t.co/HTULde1ryV: 1  
https://t.co/xPj7iG4UNd: 1  
https://t.co/sdU8lRczto: 1  
https://t.co/3pMDuLbKZF: 1  
https://t.co/tRJ3PpH2Mp: 1  
https://t.co/VI1LduMZev: 1  
https://t.co/jaaDIIdmWak: 1  
https://t.co/rPzZ3P1z1g: 1  
https://t.co/X0voYJT8kv: 1  
https://t.co/d9cZAcjrc0: 1  
https://t.co/ThFJWa660A: 1  
https://t.co/HMK34Fjukq: 1  
https://t.co/Z9qAur9Ypl: 1  
https://t.co/BKIFhtXUkX: 1  
https://t.co/SNkyaxDSv0: 1  
https://t.co/1LscnrhLOY: 1  
https://t.co/RFQ85QLIK9: 1  
https://t.co/5W4hD2f7vD: 1  
https://t.co/8V5tBc9hnc: 1  
https://t.co/Ra80rxS9wN: 1  
https://t.co/NgTrKnNKYR: 1  
https://t.co/wEvoB2YCEr: 1
```


Google drive link

<https://drive.google.com/open?id=13dJ2lr0RD6n63UQkkeFzBLgjzzbnfUv>