BDT Final Project

Facebook Post Analysis

Professor : Mrudula Mukadam
Student : Ling Sun 986265

Facebook Post Analysis - inspirations

1. For a single account, What the most common words used in all of the posts
2. For a single account, how many posts posted each year?
3. For a single account, what are the most commented posts?
4. What are the most liked posts?
5. Who are my ‘best friends’?
6. ……

Development Environment

OS : mac OS high Sierra Version 10.13.4

Hadoop 2.6.5

Apache Spark 2.6

Hive 2.3.1

Python 3.6.4

Kafka 1.1.0

Problems

1. Failed to start Hadoop server
2. Failed to start namenode server
3. Failed to start hive metastore server

But till last Sunday, I got the dev environment set

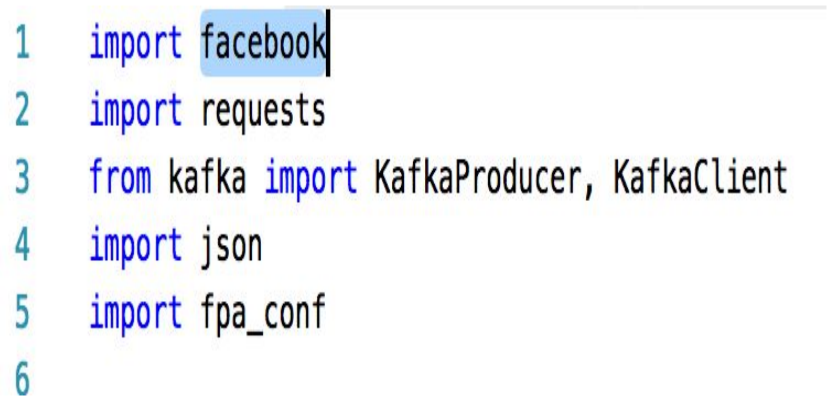
4. Spark didn’t connect with Hive successfully
5. Don’t know anything about Python

……there will be more to be found

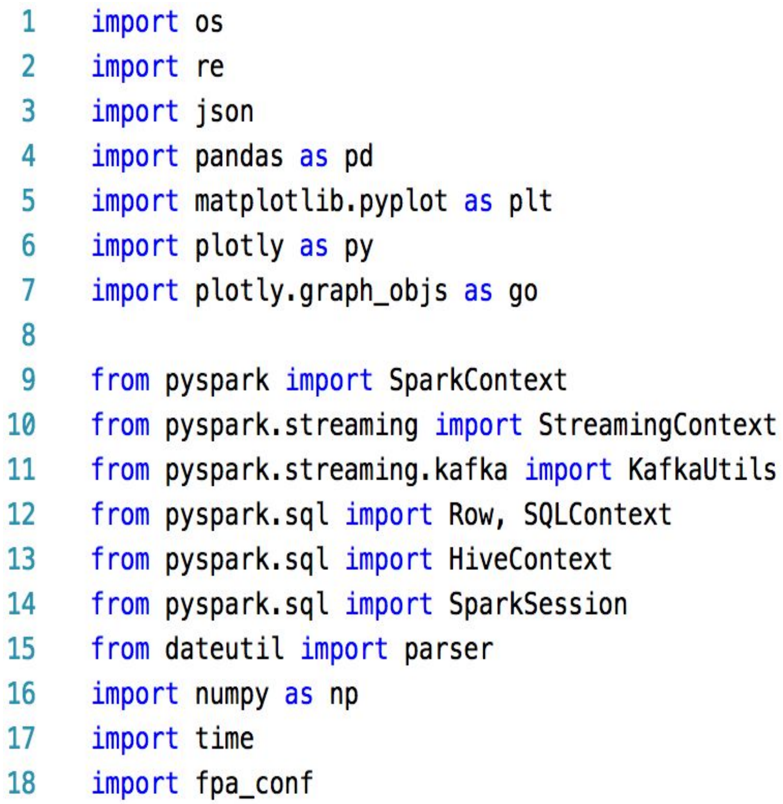
Data Flow

1. Using Facebook Graph API and Http to request real-time streaming data
2. Receiving and parsing data
3. Sending data to Kafka
4. Using Spark Streaming to read data from Kafka
5. Saving data to HDFS via Hive
6. Using Spark SQL to query and analyse
7. Using Plotly to visualize data

Code - Libs For Python

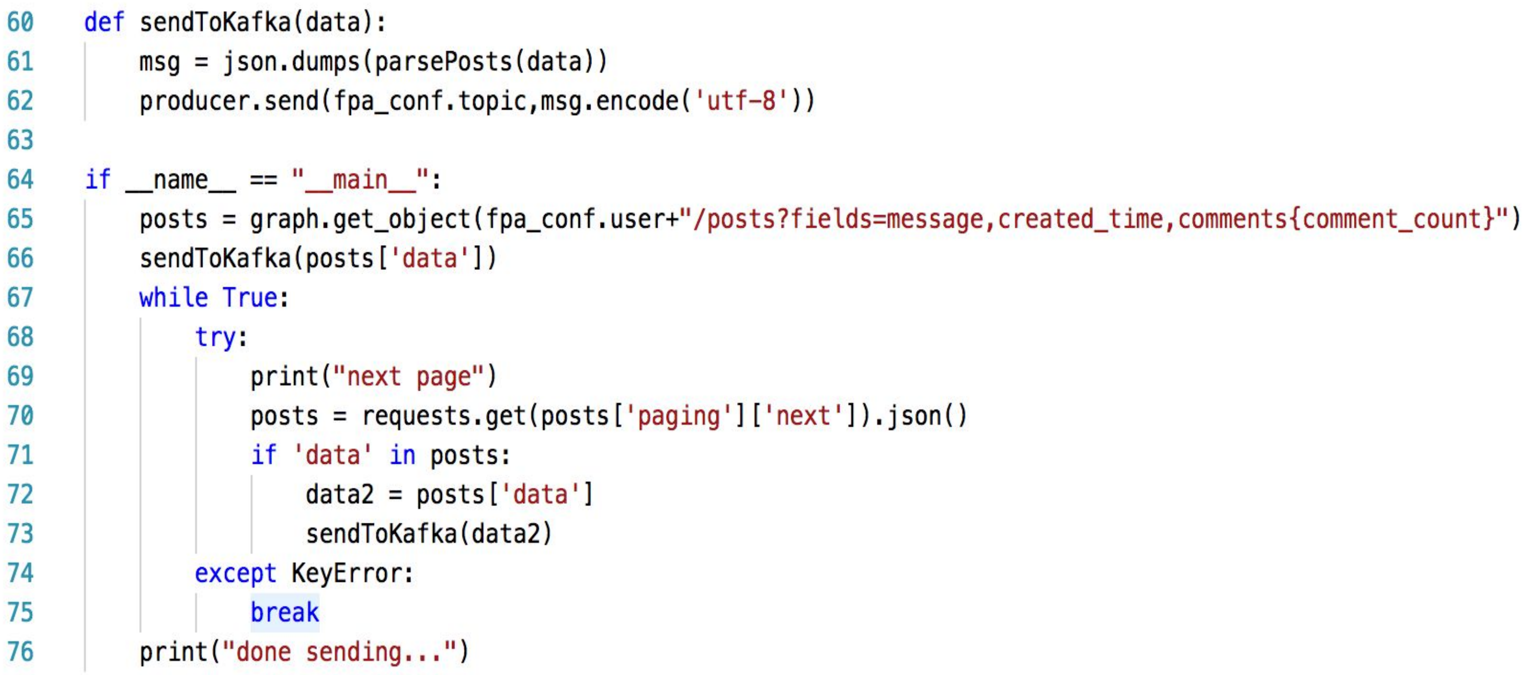


```python
import facebook
import requests
from kafka import KafkaProducer, KafkaClient
import json
import fpa_conf

```



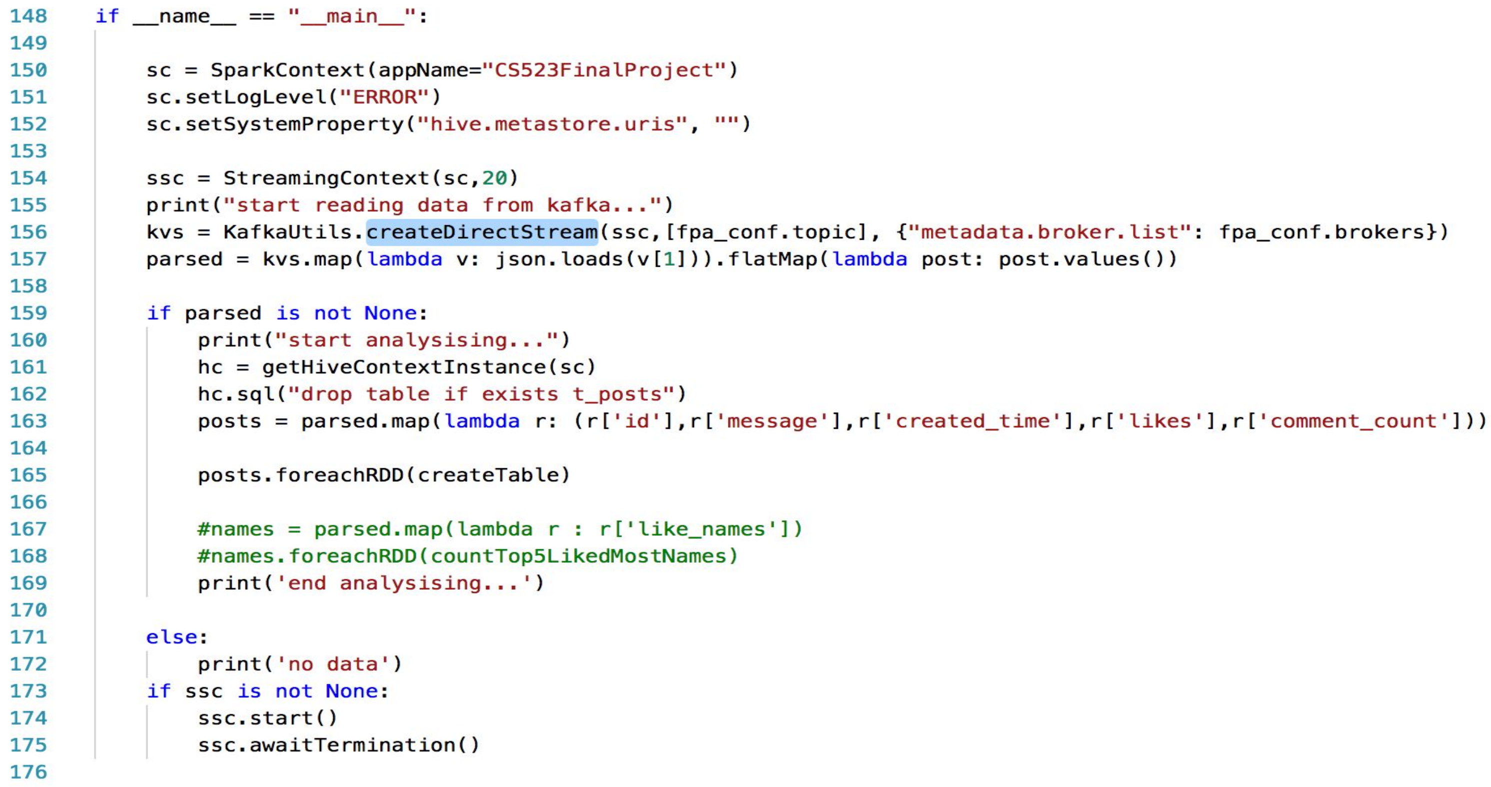
```python
import os
import re
import json
import pandas as pd
import matplotlib.pyplot as plt
import plotly as py
import plotly.graph_objs as go

from pyspark import SparkContext
from pyspark.streaming import StreamingContext
from pyspark.streaming.kafka import KafkaUtils
from pyspark.sql import Row, SQLContext
from pyspark.sql import HiveContext
from pyspark.sql import SparkSession
from dateutil import parser
import numpy as np
import time
import fpa_conf
```

Code - Producer



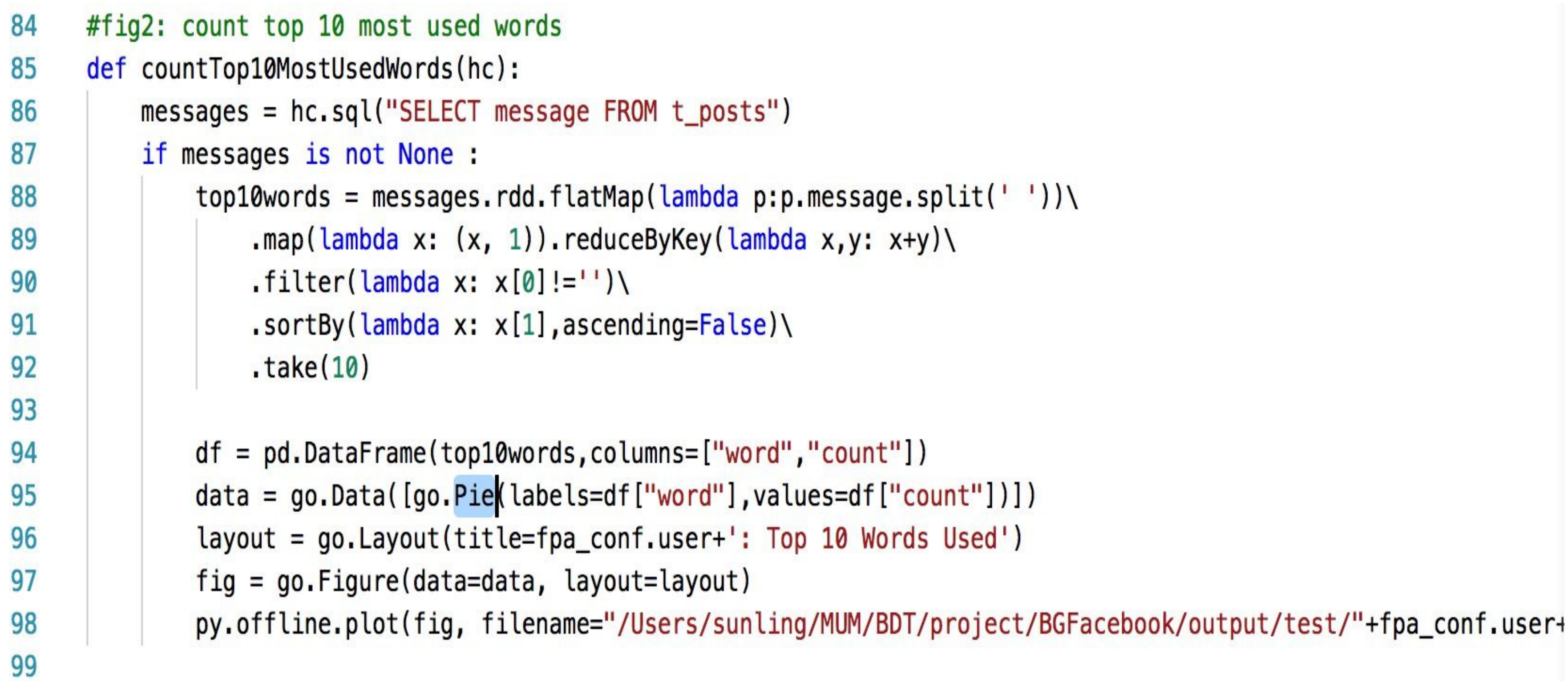
```python
def sendToKafka(data):
    msg = json.dumps(parsePosts(data))
    producer.send(fpa_conf.topic,msg.encode('utf-8'))

if __name__ == "__main__":
    posts = graph.get_object(fpa_conf.user+"/posts?fields=message,created_time,comments{comment_count}")
    sendToKafka(posts['data'])
    while True:
        try:
            print("next page")
            posts = requests.get(posts['paging']['next']).json()
            if 'data' in posts:
                data2 = posts['data']
                sendToKafka(data2)
        except KeyError:
            break
    print("done sending...")
```

Code - Consumer



```
if __name__ == "__main__":

    sc = SparkContext(appName="CS523FinalProject")
    sc.setLogLevel("ERROR")
    sc.setSystemProperty("hive.metastore.uris", "")

    ssc = StreamingContext(sc,20)
    print("start reading data from kafka...")
    kvs = KafkaUtils.createDirectStream(ssc,[fpa_conf.topic], {"metadata.broker.list": fpa_conf.brokers})
    parsed = kvs.map(lambda v: json.loads(v[1])).flatMap(lambda post: post.values())

    if parsed is not None:
        print("start analysising...")
        hc = getHiveContextInstance(sc)
        hc.sql("drop table if exists t_posts")
        posts = parsed.map(lambda r: (r['id'],r['message'],r['created_time'],r['likes'],r['comment_count']))

        posts.foreachRDD(createTable)

        #names = parsed.map(lambda r : r['like_names'])
        #names.foreachRDD(countTop5LikedMostNames)
        print('end analysising...')

    else:
        print('no data')
    if ssc is not None:
        ssc.start()
        ssc.awaitTermination()

```

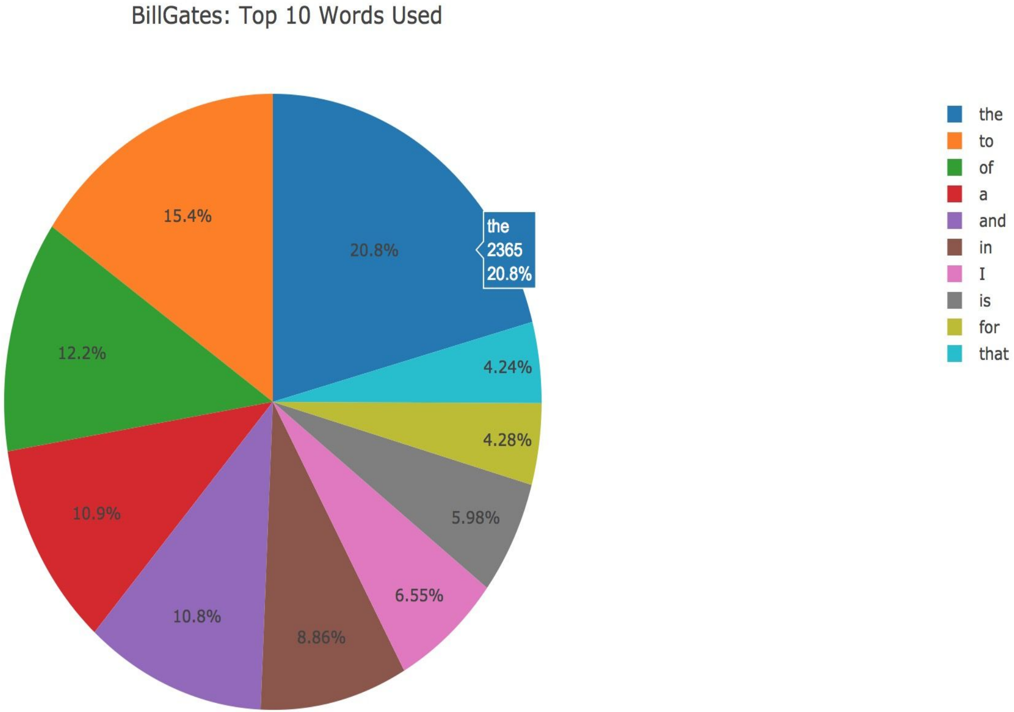
Code - Consumer



```
#fig2: count top 10 most used words
def countTop10MostUsedWords(hc):
    messages = hc.sql("SELECT message FROM t_posts")
    if messages is not None :
        top10words = messages.rdd.flatMap(lambda p:p.message.split(' '))\
            .map(lambda x: (x, 1)).reduceByKey(lambda x,y: x+y)\
            .filter(lambda x: x[0]!='')\
            .sortBy(lambda x: x[1],ascending=False)\
            .take(10)

        df = pd.DataFrame(top10words,columns=["word","count"])
        data = go.Data([go.Pie(labels=df["word"],values=df["count"])])
        layout = go.Layout(title=fpa_conf.user+': Top 10 Words Used')
        fig = go.Figure(data=data, layout=layout)
        py.offline.plot(fig, filename="/Users/sunling/MUM/BDT/project/BGFacebook/output/test/"+fpa_conf.user+
```

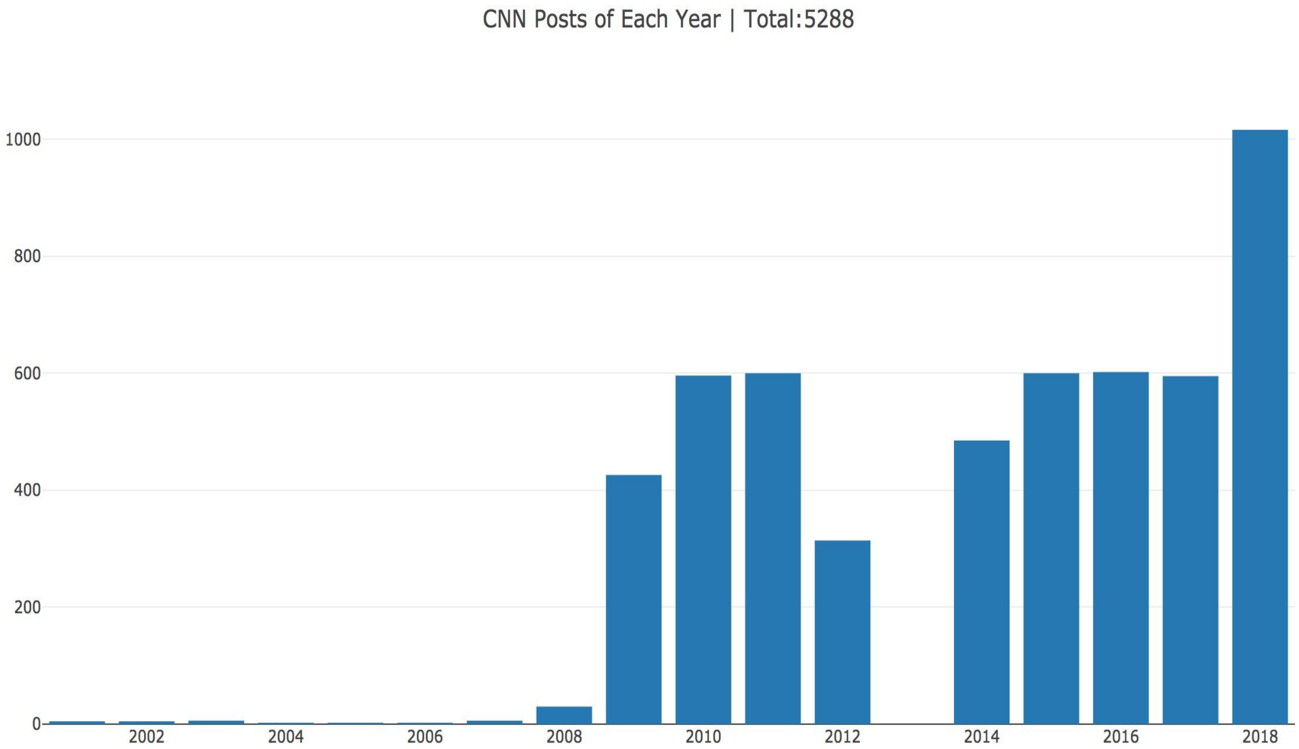
Outputs - Pie

https://www.facebook.com/BillGates



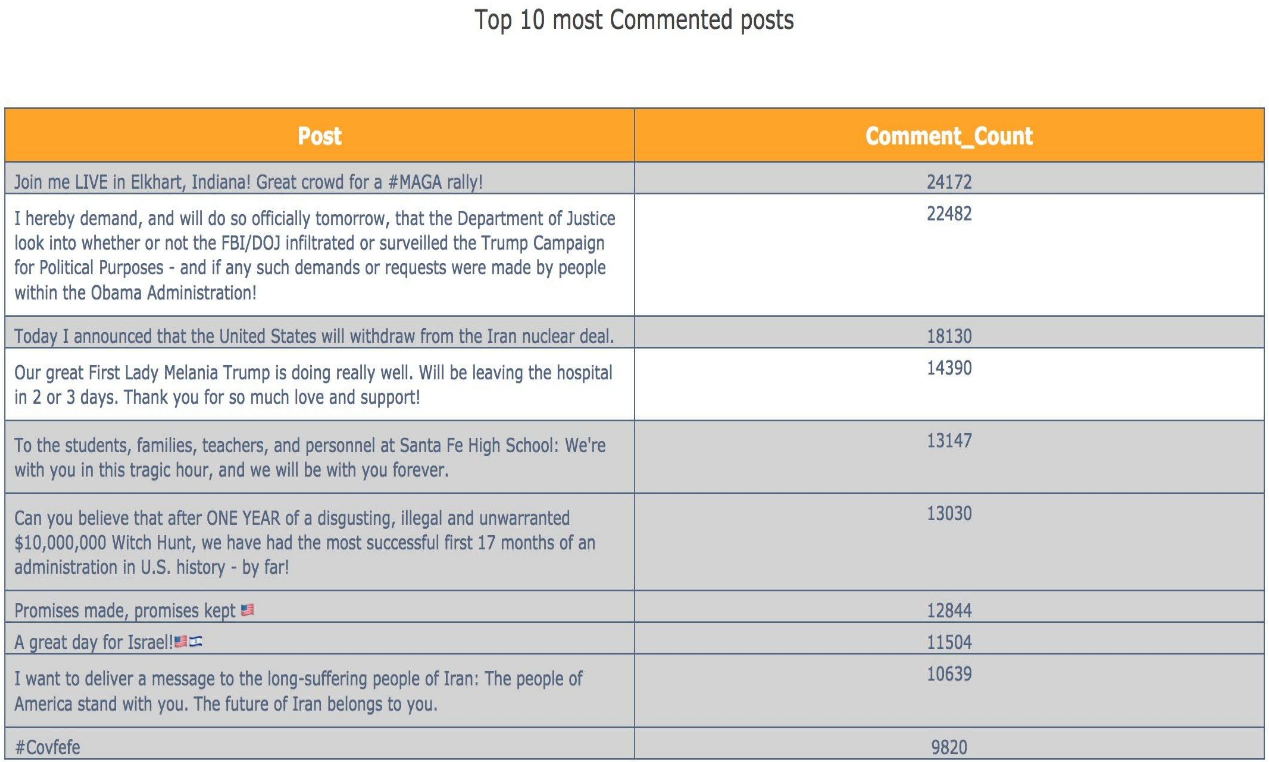
Outputs - Bar

https://www.facebook.com/cnn/



Outputs - Table

https://www.facebook.com/DonaldTrump/



Top 10 most Commented posts

Post	Comment_Count
Join me LIVE in Elkhart, Indiana! Great crowd for a #MAGA rally!	24172
I hereby demand, and will do so officially tomorrow, that the Department of Justice look into whether or not the FBI/DOJ infiltrated or surveilled the Trump Campaign for Political Purposes - and if any such demands or requests were made by people within the Obama Administration!	22482
Today I announced that the United States will withdraw from the Iran nuclear deal.	18130
Our great First Lady Melania Trump is doing really well. Will be leaving the hospital in 2 or 3 days. Thank you for so much love and support!	14390
To the students, families, teachers, and personnel at Santa Fe High School: We're with you in this tragic hour, and we will be with you forever.	13147
Can you believe that after ONE YEAR of a disgusting, illegal and unwarranted $10,000,000 Witch Hunt, we have had the most successful first 17 months of an administration in U.S. history - by far!	13030
Promises made, promises kept 🇺🇸	12844
A great day for Israel!🇺🇸🇮🇱	11504
I want to deliver a message to the long-suffering people of Iran: The people of America stand with you. The future of Iran belongs to you.	10639
#Covfefe	9820

Demo

```
#start hadoop

cd /usr/local/Cellar/hadoop-2.6.5/

sbin/start-dfs.sh

sbin/start-yarn.sh

#start hive metastore

hive --service metastore

#start Kafka, mysql, spark

brew service kafka start

......
```

```
# run consumer

spark-submit --jars /usr/local/Cellar/kafka/1.1.0/libexec/libs/spark-streaming-kafka-assembly_2.11-1.6.1.jar /Users/sunling/MUM/BDT/project/BGFacebook/BGConsumer.py

# run producer

python /Users/sunling/MUM/BDT/project/BGFacebook/BGProducer.py
```

Thank you!

Any questions ?