Spark Streaming: Ejemplo con Twint [1/2]

Usaremos en nuestro caso twint para obtener la información:

https://github.com/twintproject/twint (https://github.com/twintproject/twint)

Instalación de la librería twint

In [1]:

```
# pip install --user --upgrade
# -e git+https://github.com/twintproject/twint.git@origin/master#egg=twint
```

In [2]:

```
# pip install nest_asyncio
```

Importar dependencias

In [3]:

```
import socket
import sys
import requests
import json
import twint
from datetime import datetime, timedelta
from time import sleep
```

In [4]:

```
import nest_asyncio
nest_asyncio.apply()

# Tweets de las 2 últimas horas
current_date = datetime.now()
current_end_date = current_date - timedelta(minutes=240)

def twint_to_pandas(columns):
    return twint.output.panda.Tweets_df[columns]

# Configuración de twint
c= twint.Config()
c.Username = "20m"
c.Since = current_end_date.strftime("%Y-%m-%d %H:%M:%S")
c.Until = current_date.strftime("%Y-%m-%d %H:%M:%S")
c.Pandas = True

# Run
twint.run.Search(c)
df_pd = twint_to_pandas(["date", "username",
                         "tweet", "hashtags",
                         "nlikes"])
```

```
1597883271622365187 2022-11-30 09:20:33 +0000 <20m> ▶ Argentina nece
sita a Messi para creer de nuevo  https://t.co/Rtill6Gi95 (https://
t.co/Rtill6Gi95)
1597882122215235587 2022-11-30 09:15:59 +0000 <20m> Convierte su vie
jo Toyota en un flamante Ferrari y es detenido por falsificación  ht
tps://t.co/nc5HqPiR1M (https://t.co/nc5HqPiR1M) El blog de @elbecari
o
1597881074939478021 2022-11-30 09:11:50 +0000 <20m> A cuántos grados
debo poner la lavadora para ahorrar un 55% de luz y que la ropa salg
a limpia  https://t.co/GD94oAfKCo (https://t.co/GD94oAfKCo)
1597880006126043137 2022-11-30 09:07:35 +0000 <20m> 🎙#Entrevista a
 Omar Montes (@omarmontesSr): "La fama me ha cambiado en el sentido
 de que cuando voy al cine, en vez de colarnos, ahora invito a todo
s. No miro el dinero. No es que antes lo mirara, pero como no lo ten
ía..."  https://t.co/RiBY67FTsQ (https://t.co/RiBY67FTsQ) Por @d_mat
eo
1597878747830575104 2022-11-30 09:02:35 +0000 <20m> Aprovechan un re
to viral de TikTok para engañar a miles de usuarios con malware  htt
ps://t.co/tvnOXF3lRc (https://t.co/tvnOXF3lRc)
1597877693349072897 2022-11-30 08:58:23 +0000 <20m> Acuerdo de divor
```

In [5]:

```
from pyspark.sql import SparkSession

spark = SparkSession.builder.appName("pandasToSparkDF").getOrCreate()

df_pd = spark.createDataFrame(df_pd)
lines = df_pd.select("tweet")
print(lines)
```

```
/usr/local/spark/python/pyspark/sql/pandas/conversion.py:474: FutureWa
rning: iteritems is deprecated and will be removed in a future versio
n. Use .items instead.
  for column, series in pdf.iteritems():
/usr/local/spark/python/pyspark/sql/pandas/conversion.py:486: FutureWa
rning: iteritems is deprecated and will be removed in a future versio
n. Use .items instead.
  for column, series in pdf.iteritems():

DataFrame[tweet: string]
```

Procesar los tweets

In [6]:

```
from pyspark.sql.functions import *
from pyspark.sql.types import *
from pyspark.sql import functions as F
import re

def preprocessing(lines):
    words = lines.na.replace("", None)
    words = lines.na.drop()
    words = lines.withColumn("tweet", F.regexp_replace("tweet", r'http\S+', ''))
    words = lines.withColumn("tweet", F.regexp_replace("tweet", '@\w+', ''))
    words = lines.withColumn("tweet", F.regexp_replace("tweet", '#', ''))
    words = lines.withColumn("tweet", F.regexp_replace("tweet", 'RT', ''))
    words = lines.withColumn("tweet", F.regexp_replace("tweet", ':', ''))
    full_tweet = words.toJSON()
    return full_tweet
```

Función para enviar los tweets al socket

In [7]:

```
def send_tweets_to_spark(data, tcp_connection):
    for row in data.collect():
        line = row + "\n"
        try:
            tcp_connection.send(line.encode("utf-8"))
        except:
            e = sys.exc_info()[0]
            print("Error connection: %s" % e)
```

Crear el socket donde se almacena la información

In [8]:

```
TCP_IP = "localhost"
TCP_PORT = 9009
conn = None

s = socket.socket(socket.AF_INET, socket.SOCK_STREAM)
s.bind((TCP_IP, TCP_PORT))
s.listen(1)
print("Waiting for TCP connection...")
conn, addr = s.accept()
print("Connected... Starting getting tweets.")
data = preprocessing(lines)
send_tweets_to_spark(data, conn)
```

```
Waiting for TCP connection...

---------------------------------------------------------------------------
-----
KeyboardInterrupt                         Traceback (most recent call
 last)
Cell In [8], line 9
      7 s.listen(1)
      8 print("Waiting for TCP connection...")
----> 9 conn, addr = s.accept()
     10 print("Connected... Starting getting tweets.")
     11 data = preprocessing(lines)

File /opt/conda/lib/python3.10/socket.py:293, in socket.accept(self)
    286 def accept(self):
    287     """accept() -> (socket object, address info)
    288
    289     Wait for an incoming connection.  Return a new socket
    290     representing the connection, and the address of the clien
t.
    291     For IP sockets, the address info is a pair (hostaddr, por
t).
    292     """
--> 293     fd, addr = self._accept()
    294     sock = socket(self.family, self.type, self.proto, fileno=f
d)
    295     # Issue #7995: if no default timeout is set and the listen
ing
    296     # socket had a (non-zero) timeout, force the new socket in
blocking
    297     # mode to override platform-specific socket flags inherita
nce.

KeyboardInterrupt:
```

In [ ]:

```

```