Final_Project_2.5

March 10, 2023

[1]:
```
import sys
print(sys.version)
```

```
3.8.15 | packaged by conda-forge | (default, Nov 22 2022, 08:46:39)
[GCC 10.4.0]
```

[2]:
```
spark.version
```

[2]:
```
'3.1.3'
```

[3]:
```
import pandas as pd
import numpy as np
pd.set_option('display.max_colwidth', None)
pd.reset_option('display.max_rows')
from itertools import compress
from pyspark.sql.functions import *
from pyspark.sql.types import *
import seaborn as sns
import matplotlib.pyplot as plt
warnings.filterwarnings(action='ignore')
```

[4]:
```
from pyspark.sql import SparkSession
from pyspark import SparkContext
from pyspark.sql import SQLContext
from pyspark.sql import Row
from pyspark.sql.functions import col
```

[5]:
```
spark.conf.set("spark.sql.repl.eagerEval.enabled",True)
```

[6]:
```
%%time
tweets_en = spark.read.parquet('gs://chen26-bdp/filtered')
```

```
CPU times: user 8.96 ms, sys: 866 µs, total: 9.82 ms
Wall time: 7.13 s
```

[6]:
```
%%time
twitter = spark.read.parquet('gs://chen26-bdp/original_data')
```

```
CPU times: user 10.3 ms, sys: 548 µs, total: 10.8 ms
Wall time: 8.35 s
```

```
23/03/04 02:58:42 WARN org.apache.spark.sql.catalyst.util.package: Truncated the
string representation of a plan since it was too large. This behavior can be
adjusted by setting 'spark.sql.debug.maxToStringFields'.
```

Exclude text that do not contain keywords

[7]:
```
keywords = ['college', 'high', 'university', 'students'
            , 'public', 'private', 'secondary', 'primary', 'education',␣
 ↪'undergraduate','graduate']
```

[8]:
```
#filter out rows that do not contain words in keywords
twitter = twitter.withColumn('lower', lower(col('text')))
filter_twitter = twitter.filter(col('lower').rlike('|'.join(keywords)))
```

[9]:
```
twitter_eng = filter_twitter.filter(col('lang') == 'en')
from pyspark.sql import functions as F
from pyspark.sql import types as t
from pyspark.sql.types import ArrayType, IntegerType, BooleanType

eng_ord=F.udf(lambda x: [ord(a) for a in x],t.ArrayType(IntegerType()))

def english_filter(x):
    for index in range(len(x)):
        if x[index] > 128:
            return False
        else:
            return True

filter_udf = F.udf(english_filter, BooleanType())
tweets = twitter_eng.filter(filter_udf(eng_ord('text')) == True)
```

[31]:
```
tweets.count()
```

[31]:
```
39621120
```

[ ]:
```

```

[13]:
```
tweets_df = tweets.select([tweets.created_at,
                           tweets.id_str.alias('tweet_id'),
```

```
                                tweets.user['id_str'].alias('user_id'),
                                tweets.user['name'].alias('user_name'),
                                tweets.user['verified'].alias('verified'),
                                tweets.user['followers_count'].
↪alias('followers_count'),
                                tweets.user['description'].
↪alias('user_description'),

                                tweets.text,
                                tweets.lang.alias('tweet_language'),

                                tweets.retweeted,
                                tweets.retweeted_from,
                                tweets.retweeted_status,
                                tweets.retweeted_status.retweet_count.
↪alias('rt_count'),
                                tweets.retweeted_status.favorite_count.
↪alias('rt_fav'),
                                tweets.retweeted_status.quote_count.
↪alias('rt_quo'),
                                tweets.retweeted_status.user['name'].
↪alias('rt_user_name')])



                                #tweets.quoted_status.alias('tweet_quote'),
                                #tweets.place.country.alias('tweet_country'),
                                #tweets.place.full_name.alias('tweet_location')])
```

[40]:
```
tweets_df.select('user_id', 'user_name', 'rt_count', 'rt_quo', 'text')
```

[40]:
```
+------------------+------------------------------+--------+------+------------
---------------+
|           user_id|                     user_name|rt_count|rt_quo|
text|
+------------------+------------------------------+--------+------+------------
---------------+
|1251367945042886658|            Andres.simonetti25|       7|     0|       RT
@toniprnews: #…|
|1101871913771692032|       NBA   …|    null|  null|#     #  https…|
|         297558650|                Dumb Hollywood|      11|     0|       RT
@baseballinpix…|
|          30237884|          Covid's Not Done …|     191|    54|       RT
@unlearn16twee…|
|        1202273905|              Nancy Nord Bence|    1913|    53|       RT
@RepRaskin: Si…|
```

```
| 791367120617607168|                 m@ri@   |    null|  null|       the
fact that I g…|
|1545658665658503168|          Bold Bottom Bro  |      34|     0|       RT
@bobbycruising…|
|1562496244034850818|                      Jodeen|     184|     0|       RT
@brianBowiexxx…|
|1279641634234175488|                   Ajdin Delic|      14|     2|       RT
@PrinceZombo2:…|
|         2613254236|                 Roseline  |    null|  null|       Selling
my 4x The…|
|1413848524517687296|          Jisoo's plushie plug|    null|  null|       If
you're confusi…|
|1591411552388890624|                 cinna ↓28lb|       2|     0|       RT
@oatmilkera: i…|
|           23242642|             Kikered / Kikirin|     526|    15|       RT
@ErinInTheMorn…|
|1498572561789206529|               Lillian Sheriff|      62|     1|       RT
@WhatAboutClas…|
|1437582409906655236|          Willie Edward Tay…|    null|  null|
@bungalow3500 Tha…|
|          957636463|                   LoneWolf |    1046|   232|       RT
@toosii2x: my …|
|1449084294828371974|            Proff Lann andrews|    null|  null|
Homework | nursin…|
|1350274870827954180|                   NormalMark|    null|  null|
Wasn't wild about…|
| 894937661709979649|                   M     |    2105|    66|       RT
@leilaclaire: …|
|1597799368262037504|                 ZaxNewsStand|    null|  null|
Former Alabama Re…|
+-------------------+-----------------------------+--------+------+------------
---------------+
only showing top 20 rows
```

Original tweets

[19]:
```
ori = tweets_df.filter(col('retweeted_status').isNull())
```

[14]:
```
o = tweets_df.filter(col('retweeted_status').getItem('retweeted').isNull())
```

[17]:
```
o.select('user_name').show()
```

```
[Stage 4:>                                                          (0 + 1) / 1]

+--------------------------------+
|                       user_name|
+--------------------------------+
|          NBA   …|
```

```
|                    m@ri@   |
|                    Roseline  |
|              Jisoo's plushie plug|
|              Willie Edward Tay…|
|                Proff Lann andrews|
|                        NormalMark|
|                      ZaxNewsStand|
|                   Kevin Kershner|
|                           slmjim|
|                 Chicago H & F TC|
|                              Paul|
|                    Tania   |
|                     Tait Howard|
|                     Kelly Greco|
|              Sequoia Nagamatsu…|
|           …|
|          NBA   …|
|                      VCU_Enforcer|
|          NBA   …|
+---------------------------------+
only showing top 20 rows
```

[ ]:

[13]:
```
ori.count()
```

[13]: 13421899

[20]:
```
#top orignial posts by user name
original = ori.groupby('user_id').agg(count('*').alias('total_posts')).
 ↪orderBy(col('total_posts').desc())
original.show(10)
```

```
[Stage 7:=================================>                    (8 + 5) / 13]
+------------------+-----------+
|           user_id|total_posts|
+------------------+-----------+
|1128225338775953408|      12520|
|         219401992|       9201|
|1508968207259869185|       7769|
|          66263683|       6809|
|1576939116230455296|       6780|
|1577029442488061953|       6609|
```

```
|1463182041147576321|       6549|
| 879496394691805184|       6337|
|1582053513357537293|       6241|
|1473922978073165834|       6178|
+-------------------+-----------+
only showing top 10 rows
```

[46]:
```
ori.filter(col('user_id') == 1128225338775953408).select('created_at',␣
↪'user_name', 'text', 'followers_count')
```

[46]:
```
+-------------------+---------------+--------------------+---------------+
|         created_at|      user_name|                text|followers_count|
+-------------------+---------------+--------------------+---------------+
|Tue Oct 18 22:29:…|hsgameupdatenews|Bethlehem vs Free…|            100|
|Tue Oct 18 22:29:…|hsgameupdatenews|Atlantic Coast vs…|            100|
|Tue Oct 18 22:30:…|hsgameupdatenews|Rockledge vs Bays…|            100|
|Tue Oct 18 22:30:…|hsgameupdatenews|Florida Christian…|            100|
|Tue Oct 18 22:31:…|hsgameupdatenews|Chiles vs Nicevil…|            100|
|Tue Oct 18 22:31:…|hsgameupdatenews|Maclay vs Pensaco…|            100|
|Tue Oct 18 22:31:…|hsgameupdatenews|Miami Country Day…|            100|
|Tue Oct 18 22:31:…|hsgameupdatenews|Har-Ber vs Fayett…|            100|
|Tue Oct 18 22:31:…|hsgameupdatenews|Archer vs Harriso…|            100|
|Tue Oct 18 22:32:…|hsgameupdatenews|Winder-Barrow vs …|            100|
|Tue Oct 18 22:32:…|hsgameupdatenews|East Coweta vs Co…|            100|
|Tue Oct 18 22:32:…|hsgameupdatenews|Monroe Area vs Wh…|            100|
|Tue Oct 18 22:32:…|hsgameupdatenews|North Cobb vs Wes…|            100|
|Tue Oct 18 22:33:…|hsgameupdatenews|Manteno vs Hersch…|            100|
|Thu Sep 08 22:01:…|hsgameupdatenews|Juarez vs Phoenix…|            101|
|Thu Sep 08 22:01:…|hsgameupdatenews|Calexico vs O'Far…|            101|
|Thu Sep 08 22:02:…|hsgameupdatenews|(#12)  McDonogh v…|            101|
|Thu Sep 08 22:03:…|hsgameupdatenews|Faith Christian A…|            101|
|Thu Sep 08 22:03:…|hsgameupdatenews|MSD vs Louisiana …|            101|
|Thu Sep 08 22:04:…|hsgameupdatenews|Mission Bay vs Sw…|            101|
+-------------------+---------------+--------------------+---------------+
only showing top 20 rows
```

[21]:
```
max_reach = ori.groupby('user_id').agg(max('followers_count').
↪alias('Max_reach'))
```

[ ]:
```
max_reach.show()
```

```
[Stage 5:=================================================>(1322 + 1) / 1323]
+-------------------+---------+
```

```
|            user_id|Max_reach|
+-------------------+---------+
|1509200257191657484|       59|
|           77894076|     1094|
|         4079602277|      308|
|           24543366|      821|
|         1418496462|      431|
|         2716730338|      441|
|         2596768190|      525|
|1589588248539660288|      162|
|1115616371746807808|      160|
|           76453806|     6327|
|1505338075223896076|      241|
|           52167324|      181|
|         2563493834|    22464|
|1307848684441350144|      214|
|         4744353975|    34068|
|1406599506716770309|        2|
| 877832878830137345|   201806|
|         1730520577|      201|
|1586147771077238784|      197|
|1610719670864166956|       96|
+-------------------+---------+
only showing top 20 rows
```

[22]:
```
name = ori.groupby('user_id').agg(max('user_name').alias('user_name'))
```

[16]:
```
name.show()
```

```
[Stage 8:>                                                          (0 + 1) / 1]
+-------------------+--------------------------------+
|            user_id|                       user_name|
+-------------------+--------------------------------+
|1101871913771692032|         NBA   …|
| 791367120617607168|                    m@ri@   |
|         2613254236|                    Roseline  |
|1413848524517687296|             Jisoo's plushie plug|
|1437582409906655236|             Willie Edward Tay…|
|1449084294828371974|               Proff Lann andrews|
|1350274870827954180|                       NormalMark|
|1597799368262037504|                     ZaxNewsStand|
|1422734857776533504|                    Kevin Kershner|
|1230812852535070720|                            slmjim|
| 830056431470641152|                 Chicago H & F TC|
|          122918354|                              Paul|
```

```
|          224977551|                     Tania   |
|          192969072|                      Tait Howard|
|          147439783|                      Kelly Greco|
|         1450384922|             Sequoia Nagamatsu…|
|1595617317626564608|           …|
|         2721511461|          NBA   …|
|1199408500801253377|                     VCU_Enforcer|
|1248887297505820678|          NBA   …|
+-------------------+---------------------------------+
only showing top 20 rows
```

[23]:
```
ori_des = ori.groupby('user_id').agg(max('user_description').
 ↪alias('description'))
```

[ ]:
```
ori_des.show()
```

```
[Stage 10:>                                                         (0 + 1) / 1]
+-------------------+--------------------+
|            user_id|         description|
+-------------------+--------------------+
|1000004370786934784|Physical Security…|
|1000015310651645953|                null|
|          100001851|                null|
|          100002314|Feed of http://PR…|
|1000027408508977152|Living My Best Bo…|
|1000044397189152768|23 / 18+ / danny …|
|         1000048436|Mom, educator, co…|
|1000050752415383552|Not a bot, just a…|
|1000052019116228609|Take my advice, I…|
|1000063758402715648|Finn Valentine of…|
|          100006378|                null|
|          100007826|Britny • she/h…|
|1000082842364215296|bas likh deta hun…|
|1000092555827019776|6'3, 180 lbs | Ar…|
|1000093829700227074|2018 North Caroli…|
|1000102907621052421|Comic Artist & 2D…|
|1000119508521095169|im worth much mor…|
|1000119950827110400|          ©scorprog |
|1000136333271199749|I  mostly post ab…|
|          100014447|Audio Video Enthu…|
+-------------------+--------------------+
only showing top 20 rows
```

[27]:
```
ori_user = original.join(name, original.user_id == name.user_id, 'inner').
↪drop(name.user_id)
ori_user_df = ori_user.join(max_reach, ori_user.user_id == max_reach.user_id,␣
↪'inner').drop(max_reach.user_id)
ori_user_df = ori_user_df.join(ori_des, ori_user.user_id == ori_des.user_id,␣
↪'inner').drop(ori_des.user_id)
```

[ ]:
```
%%time
ori_user_df.write.format("parquet").\
mode('overwrite').\
save('gs://chen26-bdp/original_users')
```

323]

CPU times: user 5.67 s, sys: 1.54 s, total: 7.21 s
Wall time: 1h 12min 12s

[29]:
```
ori_user_df = spark.read.parquet('gs://chen26-bdp/original_users')
```

[33]:
```
ori_user_df
```

[33]: DataFrame[total_posts: bigint, user_id: string, user_name: string, Max_reach: bigint, description: string]

[34]:
```
ori_user_df.orderBy(col('total_posts').desc()).show()
```

[Stage 23:==> (4 + 1) / 5]

```
+-----------+------------------+-------------------+---------+----------------
----+
|total_posts|           user_id|          user_name|Max_reach|
description|
+-----------+------------------+-------------------+---------+----------------
----+
|      12520|1128225338775953408|    hsgameupdatenews|      218| hs game update
news|
|       9201|         219401992|      Dennis Stemmle|     3804|Founder -
College…|
|       7769|1508968207259869185|               JEAN|      353|     …|
|       6809|          66263683|    NJSchoolJobs.com|     4252|The leading
adver…|
|       6780|1576939116230455296|               dini|       84|
null|
|       6609|1577029442488061953|          wini teri|       37|
null|
|       6549|1463182041147576321|             sport99|       95|Welcom TV
listing…|
|       6337| 879496394691805184|    EssayPaperUK.com|      803|I AM The
```

```
Documentary|
|       6241|1582053513357537293|ilani sadiza qiop…|       18|
null|
|       6178|1473922978073165834|      Study in Naija|       48|Find All Private
…|
|       6039|1422259384525090818|       LUV 22 SPORTS|      139|Welcom TV
listing…|
|       5823|1338108136561856517|       adeliasari033|       25|   -  …|
|       5776|1214364378150998019|Every Love Live! …|     4919|
Post Ev…|
|       5658|1597301580914601985|     lioujin sopalqy|       63|
null|
|       5624|           73707872|India Education D…|     1213|One Stop
Platform…|
|       5289|1329639203144011776|              robson|       35|  HIgh School
Sports|
|       5263|          901030512|          KOINsports|      172|
null|
|       5209|1472159008886870028|          Chuo Kikuu|       97|Find All
Universi…|
|       5162|         2203526911|   X&O Coaching Jobs|     5954|College & high
sc…|
|       5139|1468164097355264002|  KQ education group|      199|THIS WEBSITE
PROV…|
+-----------+-------------------+--------------------+---------+---------------
----+
only showing top 20 rows
```

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[19]:
```
ori_max = original.join(max_reach, original.user_id == max_reach.user_id,␣
↪'inner').drop(max_reach.user_id)
```

```
[Stage 18:======>      (640 + 8) / 1323][Stage 19:======>      (650 + 8) / 1323]
```

[21]:
```
ori_posts = ori_max.join(name, ori_max.user_id == name.user_id, 'inner').
 ↪drop(name.user_id)
```

```
[Stage 18:======>      (695 + 8) / 1323][Stage 19:======>      (706 + 8) / 1323]
```

[ ]:
```
%%time
ori_posts.write.format("parquet").\
mode('overwrite').\
save('gs://chen26-bdp/original')
```

```
Exception in thread "serve-GetRows" java.net.SocketTimeoutException: Accept
timed out
        at java.net.PlainSocketImpl.socketAccept(Native Method)
        at
java.net.AbstractPlainSocketImpl.accept(AbstractPlainSocketImpl.java:409)
        at java.net.ServerSocket.implAccept(ServerSocket.java:560)
        at java.net.ServerSocket.accept(ServerSocket.java:528)
        at org.apache.spark.security.SocketAuthServer$$anon$1.run(SocketAuthServ
er.scala:64)
323]
```

```
CPU times: user 5.31 s, sys: 1.51 s, total: 6.82 s
Wall time: 58min 26s
```

[9]:
```
%%time
ori_posts = spark.read.parquet('gs://chen26-bdp/original')
```

```
CPU times: user 5.72 ms, sys: 248 µs, total: 5.97 ms
Wall time: 461 ms
```

[25]:
```
ori_posts.show()
```

```
+-----------+------------------+---------+------------------------------+
|total_posts|           user_id|Max_reach|                     user_name|
+-----------+------------------+---------+------------------------------+
|          2|1000000995814879232|       73|                     Vengeance!|
|          1|1000001158130388992|      754|          Dr. Madeline Steiner|
|          3|1000001514604318720|      751|                        shonie |
|          1|1000007184141438976|        0|                      Julianna|
|          1|         1000022436|      211|                 Alex Gauthier|
|          1|1000029845470969862|      116|             NameCannotbeBlank|
|          1|1000039793974165505|      132|                        Gina  |
|          2|1000052902457757696|     5853|                sir_Nerville |
|          1|1000063021215956992|      157|                YaGurl2Sweet |
|         10|1000068751809699840|     1130|               Veer Foundation|
|         16|1000068780582563841|      181|          The Trash God  S…|
|          1|1000073533471543296|     1213|                        ella  |
|         41|1000079474917105666|       12|       NBA   …|
```

```
|          4|1000091021668245504|       84|                  WJNoltemeyer|
|          1|         1000116332|    56978|                        Heaven|
|          2|1000117892640329730|      266|                  molly  |
|          2|1000119909932830720|     2555|                 maria  s  |
|          3|1000126711999188992|      193|                           |
|          1|1000138379038789632|      748|                   FEEQ   |
|          1|1000139068297043968|       79|                        Aayush|
+-----------+-------------------+---------+------------------------------+
only showing top 20 rows
```

[26]:
```
ori_posts.orderBy(col('total_posts').desc()).show()
```

```
[Stage 37:==============================>                          (1 + 1) / 2]
+-----------+-------------------+---------+--------------------+
|total_posts|            user_id|Max_reach|           user_name|
+-----------+-------------------+---------+--------------------+
|      12520|1128225338775953408|      218|    hsgameupdatenews|
|       9201|          219401992|     3804|      Dennis Stemmle|
|       7769|1508968207259869185|      353|                JEAN|
|       6809|           66263683|     4252|    NJSchoolJobs.com|
|       6780|1576939116230455296|       84|                dini|
|       6609|1577029442488061953|       37|           wini teri|
|       6549|1463182041147576321|       95|             sport99|
|       6337| 879496394691805184|      803|    EssayPaperUK.com|
|       6241|1582053513357537293|       18|ilani sadiza qiop…|
|       6178|1473922978073165834|       48|      Study in Naija|
|       6039|1422259384525090818|      139|       LUV 22 SPORTS|
|       5823|1338108136561856517|       25|       adeliasari033|
|       5776|1214364378150998019|     4919|Every Love Live! …|
|       5658|1597301580914601985|       63|     lioujin sopalqy|
|       5624|           73707872|     1213|India Education D…|
|       5289|1329639203144011776|       35|              robson|
|       5263|          901030512|      172|          KOINsports|
|       5209|1472159008886870028|       97|          Chuo Kikuu|
|       5162|         2203526911|     5954|   X&O Coaching Jobs|
|       5139|1468164097355264002|      199|  KQ education group|
+-----------+-------------------+---------+--------------------+
only showing top 20 rows
```

Retweets

[11]:
```
retweets = tweets_df.filter(col('retweeted_status').isNotNull())
```

[12]:
```
rt = retweets.select('user_id', 'user_name', 'retweeted_from', 'rt_count',␣
↪'rt_quo', 'text')
```

[14]:
```
rt.filter(col('retweeted_from') == 'RepRaskin')
```

[14]:
```
+------------------+------------------+-------------+--------+------+-------
-------------+
|           user_id|         user_name|retweeted_from|rt_count|rt_quo|
text|
+------------------+------------------+-------------+--------+------+-------
-------------+
|         1202273905|    Nancy Nord Bence|     RepRaskin|    1913|    53|RT
@RepRaskin: Si…|
|1435356140485988355|              PamO|     RepRaskin|    1916|    53|RT
@RepRaskin: Si…|
|          950426113|               JBE|     RepRaskin|    1923|    55|RT
@RepRaskin: Si…|
|           18375327|       Everett Will|     RepRaskin|    1924|    55|RT
@RepRaskin: Si…|
| 871591591772860416|Lauri Turner ?…|     RepRaskin|    1927|    55|RT
@RepRaskin: Si…|
| 785272833895268356|              Nesha|     RepRaskin|    1928|    55|RT
@RepRaskin: Si…|
|1558111604301934593|       Brent Morrow|     RepRaskin|    1931|    55|RT
@RepRaskin: Si…|
|1220264453482041344|   cheryl harrington|     RepRaskin|    1934|    55|RT
@RepRaskin: Si…|
|1247918255223853058|      …|     RepRaskin|    1935|    55|RT
@RepRaskin: Si…|
|           15181912|Interfaith Alliance|     RepRaskin|    1936|    55|RT
@RepRaskin: Si…|
|          384155025|         Annapolis57|     RepRaskin|    1937|    55|RT
@RepRaskin: Si…|
|         4841179036|       Harvey Lawson|     RepRaskin|    1939|    55|RT
@RepRaskin: Si…|
| 766017074745450497|             Lisa T|     RepRaskin|    1940|    55|RT
@RepRaskin: Si…|
|          234458980|       Judy Sarasohn|     RepRaskin|    1944|    55|RT
@RepRaskin: Si…|
|          798763064|   I could be anyone|     RepRaskin|    1945|    56|RT
@RepRaskin: Si…|
|           43120963|         AlpineLakes|     RepRaskin|     699|    19|RT
@RepRaskin: I …|
|           40331960|        SupportBones|     RepRaskin|     700|    20|RT
@RepRaskin: I …|
```

```
|1487836081437519875|    Cynthia D Berger|     RepRaskin|     704|    20|RT
@RepRaskin: I …|
| 910864750757666818|             Abuelis|     RepRaskin|     706|    20|RT
@RepRaskin: I …|
|1296939073773416454|Abuelita Chell …|     RepRaskin|     707|    20|RT
@RepRaskin: I …|
+------------------+--------------------+--------------+--------+------+-------
-------------+
only showing top 20 rows
```

[13]:
```
rt_count_text = rt.groupby('text').agg(max('rt_count').alias('retweet_count'))
```

[14]:
```
rt_user = rt_count_text.join(rt, rt_count_text.text == rt.text, 'left')\
.select(rt_count_text.text, rt_count_text.retweet_count, 'retweeted_from')
```

[15]:
```
rt_u = rt_user.dropDuplicates()
```

[16]:
```
rt_u_sum = rt_u.groupby('retweeted_from').agg(sum('retweet_count').
↪alias('total_retweet_count'))
```

[ ]:
```
%%time
rt_u_sum.write.format("parquet").\
mode('overwrite').\
save('gs://chen26-bdp/retweet_user_sum')
```

```
CPU times: user 2.66 s, sys: 602 ms, total: 3.26 s
Wall time: 30min 14s
```

[35]:
```
rt_u_sum = spark.read.parquet('gs://chen26-bdp/retweet_user_sum')
```

[8]:
```
rt_u_sum.orderBy(col('total_retweet_count').desc())
```

[8]:
```
+---------------+-------------------+
| retweeted_from|total_retweet_count|
+---------------+-------------------+
|         nickjr|             516855|
|   libsoftiktok|             482067|
|OccupyDemocrats|             414295|
|realDonaldTrump|             374995|
|  NasimiShabnam|             293392|
|   NoLieWithBTC|             291413|
|      _SJPeace_|             284839|
|   SportsCenter|             250517|
|   urlocalnyguy|             238989|
```

```
|          aambxt|              233700|
|     BarackObama|              220781|
|             AOC|              214729|
|         leebyyy|              208521|
|    davidhogg111|              206630|
|     Phil_Lewis_|              197817|
| thegreatkhalid|              195521|
| Public_Citizen|              193203|
|         JonWTOL|              185946|
| ColIegeStudent|              181667|
|  realchrisrufo|              181261|
+---------------+--------------------+
only showing top 20 rows
```

[36]:
```
inf = ori_user_df.join(rt_u_sum, ori_user_df.user_name == rt_u_sum.
↪retweeted_from, 'inner').drop(rt_u_sum.retweeted_from)
```

[37]:
```
inf.write.format("parquet").\
mode('overwrite').\
save('gs://chen26-bdp/influencers')
```

[38]:
```
influencers = spark.read.parquet('gs://chen26-bdp/influencers')
```

[42]:
```
influencers.orderBy(col('Max_reach').desc())
```

[42]:
```
+-----------+----------+------------+---------+--------------------+------------
-------+
|total_posts|   user_id|   user_name|Max_reach|
description|total_retweet_count|
+-----------+----------+------------+---------+--------------------+------------
-------+
|          1|  10228272|     YouTube| 77788673| like and subscribe.|
107|
|          3|  11348282|        NASA| 64878725|There's space for…|
1772|
|        158|    759251|         CNN| 61179807|It’s our job to #…|
95240|
|          3|  19923144|         NBA| 42052619|The 2022-23 NBA s…|
24773|
|         10|  26257166|SportsCenter| 41400369|Download the ESPN…|
250517|
```

```
|         19|  19426551|         NFL| 32561922|we’re baaaaaaaack…|
13113|
|          1|  10671602| PlayStation| 27836375|Official Twitter …|
111|
|        151|   1652541|     Reuters| 25730638|Top and breaking …|
23858|
|         24|  14293310|        TIME| 19450024|News and current …|
19033|
|        370|  91478624|      Forbes| 18711065|Official account …|
7778|
|          1|   2367911|         MTV| 17265417|0  days without …|
332|
|         11|   9695312|   billboard| 14087884|music • charts • …|
1136|
|          1|2455740283|     MrBeast| 13958152|I want to make th…|
7842|
|         57|  95023423|   UberFacts| 13685197|The most unimport…|
35839|
|          2|   7517222|         WWE| 12312951|The official Twit…|
1714|
|          1|  74286565|   Microsoft| 12250352|We're on a missio…|
168|
|          1|  43192807|       Pixar| 12017480|The official Twit…|
2384|
|          3|  20567939|    MeekMill| 11487511|I just wanna see …|
13892|
|          1|  17696167|    Ludacris| 11414067|http://Kidnation.com|
79|
|         62|  14511951|    HuffPost| 11370026|At HuffPost, we p…|
5812|
+-----------+----------+------------+---------+--------------------+------------
-------+
only showing top 20 rows
```

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[ ]:

[21]:
```
influencer = ori_posts.join(ori_description, ori_posts.user_id ==␣
↪ori_description.user_id, 'inner').drop(ori_description.user_id)
```

[22]:
```
%%time
influencer.write.format("parquet").\
mode('overwrite').\
save('gs://chen26-bdp/inf')
```

```
CPU times: user 1.57 s, sys: 413 ms, total: 1.99 s
Wall time: 18min 25s
```

[23]:
```
inf = spark.read.parquet('gs://chen26-bdp/inf')
```

[25]:
```
user_data = inf.join(rt_u_sum, inf.user_name == rt_u_sum.retweeted_from,␣
↪'inner').drop(rt_u_sum.retweeted_from)
```

[26]:
```
%%time
user_data.write.format("parquet").\
mode('overwrite').\
save('gs://chen26-bdp/twitter_influencer_data')
```

```
CPU times: user 14.3 ms, sys: 2.87 ms, total: 17.2 ms
Wall time: 17.7 s
```

[27]:
```
user_data = spark.read.parquet('gs://chen26-bdp/twitter_influencer_data')
```

[28]:
```
user_data
```

[28]:
```
+-----------+--------+---------+---------------+--------------------+-----------
--------+
|total_posts| user_id|Max_reach|      user_name|
description|total_retweet_count|
+-----------+--------+---------+---------------+--------------------+-----------
--------+
|          1|  659883|     1474|     desmonator|    Sweary Snarkster|
1|
|          2| 7072122|      367|         tuttut|wishing to travel…|
4|
|          4| 7492812|      372|         Mxamus|"   10   …|
4|
|         21| 9677372|    95678|           KQED|The Bay Area's @N…|
39|
```

```
|          20|13393052|  2007649|           ACLU|The ACLU is a non…|
54346|
|           1|13965002|      520|     mattmaison|  gentleman polymath|
1|
|           4|14353346|      250|      eriksmith|A proud native of…|
1|
|           1|14538937|     3220|   starrandmona|Retired from AK &…|
1|
|           1|14573504|      381|   irishcharger|NAB Coronado Ex-N…|
1|
|          65|14637243|    89966|     GoVolsXtra|University of Ten…|
156|
|           1|14713787|   627924|     CanadianPM|Official account …|
52|
|           6|15146558|   464221|       keithlaw|Baseball writer, …|
54|
|           1|15446566|     2060|           OMRF|Discoveries that …|
1|
|           4|15573372|      124|      karaann29|Boy mom. Lover of…|
36|
|           3|15592608|     6588|      PriyaRaju|Books. History. A…|
10|
|          11|15598639|    58248|GeorgiaSouthern|The official Geor…|
35|
|           7|15784401|     4320| sarajeanparker|Started using Twi…|
7|
|           1|15930919|    11836|      Spinwatch|Public interest r…|
5|
|          39|15970784|    10505|        anarcho|    People's Advisor|
45|
|           1|15971696|      163|        gdsmith|Guitarist, artist…|
10|
+------------+--------+---------+---------------+--------------------+-----------
--------+
only showing top 20 rows
```

[ ]:

[ ]: