EDA

Read Subset Data

In [6]:
```
# !hadoop fs -ls "gs://msca-bdp-tweets/final_project"
# gs://msca-bdp-tweets/final_project/part-00000-f654a635-796b-4190-88ae-6c2ee7e6f3a3-c000.json
```

In [1]:
```
from pyspark.sql.functions import *
```

In [2]:
```
data = spark.read.json('gs://msca-bdp-tweets/final_project/part-00000-f654a635-796b-4190-88ae-6c2ee7e6f3a3-c000.j
```

```
22/11/30 18:29:19 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'.
```

In [5]:
```
data.select('text').limit(5).collect()
```

Out[5]:
```
[Row(text="RT @SimuLiu: Wouldn't it be great if in high school instead of the quadratic equation I had learned to
love myself instead"),
 Row(text='RT @toyinomotoso: If you possess a valuable skill -- coding, videography, dancing, cooking, selling, r
aising capital etc - you can build an…'),
 Row(text='RT @firepitstories: College golfer Ian Johnston lost a brother and a mother to substance abuse.\n\nEve
r since, he has carried out both of the…'),
 Row(text='@kessiah44 @SenWarren Most descendants of slaves don’t go to college so they don’t have student debt.'
),
 Row(text='RT @alphajkoo: dance around me, swan — #jikookau \n\nas a boxer, jungkook is feared by many. but they
don’t know his only one secret— his obs…')]
```

In [7]:
```
data.select('created_at').limit(5).collect()
```

Out[7]:
```
[Row(created_at='Sun Oct 09 23:32:52 +0000 2022'),
 Row(created_at='Sun Oct 09 23:32:53 +0000 2022'),
 Row(created_at='Sun Oct 09 23:32:54 +0000 2022'),
 Row(created_at='Sun Oct 09 23:32:54 +0000 2022'),
 Row(created_at='Sun Oct 09 23:32:55 +0000 2022')]
```

In [30]:
```
data.select('user').limit(5).collect()
```

Out[30]:
```
[Row(user=Row(contributors_enabled=False, created_at='Tue Dec 19 07:37:20 +0000 2017', default_profile=False, def
ault_profile_image=False, description='The princess saves herself in this one', favourites_count=25052, followers
_count=365, friends_count=280, geo_enabled=True, id=943022421279432704, id_str='943022421279432704', is_translato
r=False, listed_count=0, location='JHB|PTA', name='Vuyo', profile_background_color='000000', profile_background_i
mage_url='http://abs.twimg.com/images/themes/theme1/bg.png', profile_background_image_url_https='https://abs.twim
g.com/images/themes/theme1/bg.png', profile_background_tile=False, profile_banner_url='https://pbs.twimg.com/prof
ile_banners/943022421279432704/1660944500', profile_image_url='http://pbs.twimg.com/profile_images/15684728562326
97859/TxdhY8go_normal.jpg', profile_image_url_https='https://pbs.twimg.com/profile_images/1568472856232697859/Txd
hY8go_normal.jpg', profile_link_color='19CF86', profile_sidebar_border_color='000000', profile_sidebar_fill_color
='000000', profile_text_color='000000', profile_use_background_image=False, protected=False, screen_name='Mvutyi'
, statuses_count=25671, translator_type='none', url=None, verified=False, withheld_in_countries=[])),
 Row(user=Row(contributors_enabled=False, created_at='Mon Feb 25 12:37:16 +0000 2019', default_profile=True, defa
ult_profile_image=False, description='Biotechnologist,singer,Copywriter, Internet Marketer', favourites_count=484
3, followers_count=133, friends_count=517, geo_enabled=False, id=1100011847481208832, id_str='1100011847481208832
', is_translator=False, listed_count=1, location='Akure, Nigeria', name='David Miracle', profile_background_color
='F5F8FA', profile_background_image_url='', profile_background_image_url_https='', profile_background_tile=False,
profile_banner_url=None, profile_image_url='http://pbs.twimg.com/profile_images/1310700117457068033/ZuNoT0jQ_norm
al.jpg', profile_image_url_https='https://pbs.twimg.com/profile_images/1310700117457068033/ZuNoT0jQ_normal.jpg',
profile_link_color='1DA1F2', profile_sidebar_border_color='C0DEED', profile_sidebar_fill_color='DDEEF6', profile_
text_color='333333', profile_use_background_image=True, protected=False, screen_name='DavidMiracle23', statuses_c
ount=878, translator_type='none', url='https://znap.link/David-Miracle', verified=False, withheld_in_countries=[]
)),
 Row(user=Row(contributors_enabled=False, created_at='Sat May 05 03:10:40 +0000 2012', default_profile=True, defa
ult_profile_image=False, description=None, favourites_count=12549, followers_count=90, friends_count=127, geo_ena
bled=True, id=571390621, id_str='571390621', is_translator=False, listed_count=4, location=None, name='NoblePeace
Prize', profile_background_color='C0DEED', profile_background_image_url='http://abs.twimg.com/images/themes/theme
1/bg.png', profile_background_image_url_https='https://abs.twimg.com/images/themes/theme1/bg.png', profile_backgr
ound_tile=False, profile_banner_url='https://pbs.twimg.com/profile_banners/571390621/1471812472', profile_image_u
rl='http://pbs.twimg.com/profile_images/767463243237298176/nui9kEOr_normal.jpg', profile_image_url_https='https:/
/pbs.twimg.com/profile_images/767463243237298176/nui9kEOr_normal.jpg', profile_link_color='1DA1F2', profile_sideb
ar_border_color='C0DEED', profile_sidebar_fill_color='DDEEF6', profile_text_color='333333', profile_use_backgroun
d_image=True, protected=False, screen_name='EganNoble', statuses_count=2928, translator_type='none', url=None, ve
rified=False, withheld_in_countries=[])),
 Row(user=Row(contributors_enabled=False, created_at='Mon Jan 01 04:38:12 +0000 2018', default_profile=True, defa
ult_profile_image=False, description='Nothin’', favourites_count=2093, followers_count=9, friends_count=72, geo_e
nabled=False, id=947688383932981249, id_str='947688383932981249', is_translator=False, listed_count=2, location=N
one, name='StoneCold', profile_background_color='F5F8FA', profile_background_image_url='', profile_background_ima
ge_url_https='', profile_background_tile=False, profile_banner_url=None, profile_image_url='http://pbs.twimg.com/
profile_images/947691231714381824/kbJ1-EAy_normal.jpg', profile_image_url_https='https://pbs.twimg.com/profile_im
ages/947691231714381824/kbJ1-EAy_normal.jpg', profile_link_color='1DA1F2', profile_sidebar_border_color='C0DEED',
profile_sidebar_fill_color='DDEEF6', profile_text_color='333333', profile_use_background_image=True, protected=Fa
lse, screen_name='LuzAzul21253805', statuses_count=2997, translator_type='none', url=None, verified=False, withhe
ld_in_countries=[])),
 Row(user=Row(contributors_enabled=False, created_at='Thu Nov 14 18:55:04 +0000 2019', default_profile=True, defa
ult_profile_image=False, description='I am you, you are me. this user loves the busan boyfriends  #BTS #ARMY ♡ #J
EONJUNGKOOK ♡ #PARKJIMIN', favourites_count=75, followers_count=234, friends_count=242, geo_enabled=False, id=119
5052497854795777, id_str='1195052497854795777', is_translator=False, listed_count=0, location='jikook land ', nam
e='»_𝐣𝐢𝐦𝐢𝐧𝐬𝐬𝐢™❀ˎˊ-”', profile_background_color='F5F8FA', profile_background_image_url='', profile_background_image_u
rl_https='', profile_background_tile=False, profile_banner_url='https://pbs.twimg.com/profile_banners/11950524978
54795777/1637211467', profile_image_url='http://pbs.twimg.com/profile_images/1461196920240492547/eGuqrE4d_normal.
jpg', profile_image_url_https='https://pbs.twimg.com/profile_images/1461196920240492547/eGuqrE4d_normal.jpg', pro
file_link_color='1DA1F2', profile_sidebar_border_color='C0DEED', profile_sidebar_fill_color='DDEEF6', profile_tex
t_color='333333', profile_use_background_image=True, protected=False, screen_name='rookie_kookie97', statuses_cou
nt=2424, translator_type='none', url=None, verified=False, withheld_in_countries=[]))]
```

In [43]:

```
f = data.select('user')
f = f.withColumn("location", F.col("user").getItem("location"))
```

In [45]:

```
f.select('location').show()
```

```
+--------------------+
|            location|
+--------------------+
|             JHB|PTA|
|      Akure, Nigeria|
|                null|
|                null|
|    jikook land |
|                null|
|Rent Free In Your...|
|  Elkridge, Maryland|
|             Atlanta|
|Newyork, United S...|
|St. John’s, Newfo...|
|             Florida|
|                null|
|                null|
|11.3493° N, 142.1...|
|                  NJ|
|    she they 18 □□|
|                null|
|    Ciudad de México|
|   Milwaukee-Chicago|
+--------------------+
only showing top 20 rows
```

In [22]:

```
f.printSchema()
```

```
root
 |-- user: struct (nullable = true)
 |    |-- contributors_enabled: boolean (nullable = true)
 |    |-- created_at: string (nullable = true)
 |    |-- default_profile: boolean (nullable = true)
 |    |-- default_profile_image: boolean (nullable = true)
 |    |-- description: string (nullable = true)
 |    |-- favourites_count: long (nullable = true)
 |    |-- followers_count: long (nullable = true)
 |    |-- friends_count: long (nullable = true)
 |    |-- geo_enabled: boolean (nullable = true)
 |    |-- id: long (nullable = true)
 |    |-- id_str: string (nullable = true)
 |    |-- is_translator: boolean (nullable = true)
 |    |-- listed_count: long (nullable = true)
 |    |-- location: string (nullable = true)
 |    |-- name: string (nullable = true)
 |    |-- profile_background_color: string (nullable = true)
 |    |-- profile_background_image_url: string (nullable = true)
 |    |-- profile_background_image_url_https: string (nullable = true)
 |    |-- profile_background_tile: boolean (nullable = true)
 |    |-- profile_banner_url: string (nullable = true)
 |    |-- profile_image_url: string (nullable = true)
 |    |-- profile_image_url_https: string (nullable = true)
 |    |-- profile_link_color: string (nullable = true)
 |    |-- profile_sidebar_border_color: string (nullable = true)
 |    |-- profile_sidebar_fill_color: string (nullable = true)
 |    |-- profile_text_color: string (nullable = true)
 |    |-- profile_use_background_image: boolean (nullable = true)
 |    |-- protected: boolean (nullable = true)
 |    |-- screen_name: string (nullable = true)
 |    |-- statuses_count: long (nullable = true)
 |    |-- translator_type: string (nullable = true)
 |    |-- url: string (nullable = true)
 |    |-- verified: boolean (nullable = true)
 |    |-- withheld_in_countries: array (nullable = true)
 |    |    |-- element: string (containsNull = true)
```

In [12]:

```
data.select('retweet_count').limit(5).collect()
```

Out[12]:

```
[Row(retweet_count=0),
 Row(retweet_count=0),
 Row(retweet_count=0),
 Row(retweet_count=0),
 Row(retweet_count=0)]
```

In [10]:

```
data.printSchema()
```

```
root
 |-- coordinates: struct (nullable = true)
 |    |-- coordinates: array (nullable = true)
 |    |    |-- element: double (containsNull = true)
 |    |-- type: string (nullable = true)
 |-- created_at: string (nullable = true)
 |-- display_text_range: array (nullable = true)
 |    |-- element: long (containsNull = true)
 |-- entities: struct (nullable = true)
 |    |-- hashtags: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |-- text: string (nullable = true)
 |    |-- media: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |    |-- description: string (nullable = true)
 |    |    |    |    |-- embeddable: boolean (nullable = true)
 |    |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |    |-- title: string (nullable = true)
 |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |-- id: long (nullable = true)
 |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |-- media_url: string (nullable = true)
 |    |    |    |-- media_url_https: string (nullable = true)
 |    |    |    |-- sizes: struct (nullable = true)
 |    |    |    |    |-- large: struct (nullable = true)
 |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |-- medium: struct (nullable = true)
 |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |-- small: struct (nullable = true)
 |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |-- thumb: struct (nullable = true)
 |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |-- source_status_id: long (nullable = true)
 |    |    |    |-- source_status_id_str: string (nullable = true)
 |    |    |    |-- source_user_id: long (nullable = true)
 |    |    |    |-- source_user_id_str: string (nullable = true)
 |    |    |    |-- type: string (nullable = true)
 |    |    |    |-- url: string (nullable = true)
 |    |-- symbols: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |-- text: string (nullable = true)
 |    |-- urls: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |-- url: string (nullable = true)
 |    |-- user_mentions: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- id: long (nullable = true)
 |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |-- name: string (nullable = true)
 |    |    |    |-- screen_name: string (nullable = true)
 |-- extended_entities: struct (nullable = true)
 |    |-- media: array (nullable = true)
 |    |    |-- element: struct (containsNull = true)
 |    |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |    |-- description: string (nullable = true)
 |    |    |    |    |-- embeddable: boolean (nullable = true)
 |    |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |    |-- title: string (nullable = true)
 |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |-- id: long (nullable = true)
 |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |-- media_url: string (nullable = true)
 |    |    |    |-- media_url_https: string (nullable = true)
 |    |    |    |-- sizes: struct (nullable = true)
 |    |    |    |    |-- large: struct (nullable = true)
 |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |-- medium: struct (nullable = true)
 |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |-- small: struct (nullable = true)
 |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |-- thumb: struct (nullable = true)
 |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |-- source_status_id: long (nullable = true)
 |    |    |    |-- source_status_id_str: string (nullable = true)
 |    |    |    |-- source_user_id: long (nullable = true)
 |    |    |    |-- source_user_id_str: string (nullable = true)
 |    |    |    |-- type: string (nullable = true)
 |    |    |    |-- url: string (nullable = true)
 |    |    |    |-- video_info: struct (nullable = true)
 |    |    |    |    |-- aspect_ratio: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- duration_millis: long (nullable = true)
 |    |    |    |    |-- variants: array (nullable = true)
 |    |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |    |-- bitrate: long (nullable = true)
 |    |    |    |    |    |    |-- content_type: string (nullable = true)
 |    |    |    |    |    |    |-- url: string (nullable = true)
 |-- extended_tweet: struct (nullable = true)
 |    |-- display_text_range: array (nullable = true)
 |    |    |-- element: long (containsNull = true)
 |    |-- entities: struct (nullable = true)
 |    |    |-- hashtags: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- indices: array (nullable = true)
```

```
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- text: string (nullable = true)
 |    |    |-- media: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |    |-- description: string (nullable = true)
 |    |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |    |-- id: long (nullable = true)
 |    |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- media_url: string (nullable = true)
 |    |    |    |    |-- media_url_https: string (nullable = true)
 |    |    |    |    |-- sizes: struct (nullable = true)
 |    |    |    |    |    |-- large: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |-- medium: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |-- small: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |-- thumb: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |-- type: string (nullable = true)
 |    |    |    |    |-- url: string (nullable = true)
 |    |    |    |    |-- video_info: struct (nullable = true)
 |    |    |    |    |    |-- aspect_ratio: array (nullable = true)
 |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |-- duration_millis: long (nullable = true)
 |    |    |    |    |    |-- variants: array (nullable = true)
 |    |    |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |    |    |-- bitrate: long (nullable = true)
 |    |    |    |    |    |    |    |-- content_type: string (nullable = true)
 |    |    |    |    |    |    |    |-- url: string (nullable = true)
 |    |    |-- symbols: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- text: string (nullable = true)
 |    |    |-- urls: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- url: string (nullable = true)
 |    |    |-- user_mentions: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- id: long (nullable = true)
 |    |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- name: string (nullable = true)
 |    |    |    |    |-- screen_name: string (nullable = true)
 |    |-- extended_entities: struct (nullable = true)
 |    |    |-- media: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |    |-- description: string (nullable = true)
 |    |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |    |-- id: long (nullable = true)
 |    |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- media_url: string (nullable = true)
 |    |    |    |    |-- media_url_https: string (nullable = true)
 |    |    |    |    |-- sizes: struct (nullable = true)
 |    |    |    |    |    |-- large: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |-- medium: struct (nullable = true)
```

```
|    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |-- small: struct (nullable = true)
|    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |-- thumb: struct (nullable = true)
|    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |-- type: string (nullable = true)
|    |    |    |    |-- url: string (nullable = true)
|    |    |    |    |-- video_info: struct (nullable = true)
|    |    |    |    |    |-- aspect_ratio: array (nullable = true)
|    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |-- duration_millis: long (nullable = true)
|    |    |    |    |    |-- variants: array (nullable = true)
|    |    |    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |    |    |-- bitrate: long (nullable = true)
|    |    |    |    |    |    |    |-- content_type: string (nullable = true)
|    |    |    |    |    |    |    |-- url: string (nullable = true)
|    |-- full_text: string (nullable = true)
|-- favorite_count: long (nullable = true)
|-- favorited: boolean (nullable = true)
|-- filter_level: string (nullable = true)
|-- geo: struct (nullable = true)
|    |-- coordinates: array (nullable = true)
|    |    |-- element: double (containsNull = true)
|    |-- type: string (nullable = true)
|-- id: long (nullable = true)
|-- id_str: string (nullable = true)
|-- in_reply_to_screen_name: string (nullable = true)
|-- in_reply_to_status_id: long (nullable = true)
|-- in_reply_to_status_id_str: string (nullable = true)
|-- in_reply_to_user_id: long (nullable = true)
|-- in_reply_to_user_id_str: string (nullable = true)
|-- is_quote_status: boolean (nullable = true)
|-- lang: string (nullable = true)
|-- place: struct (nullable = true)
|    |-- bounding_box: struct (nullable = true)
|    |    |-- coordinates: array (nullable = true)
|    |    |    |-- element: array (containsNull = true)
|    |    |    |    |-- element: array (containsNull = true)
|    |    |    |    |    |-- element: double (containsNull = true)
|    |    |-- type: string (nullable = true)
|    |-- country: string (nullable = true)
|    |-- country_code: string (nullable = true)
|    |-- full_name: string (nullable = true)
|    |-- id: string (nullable = true)
|    |-- name: string (nullable = true)
|    |-- place_type: string (nullable = true)
|    |-- url: string (nullable = true)
|-- possibly_sensitive: boolean (nullable = true)
|-- quote_count: long (nullable = true)
|-- quoted_status: struct (nullable = true)
|    |-- created_at: string (nullable = true)
|    |-- display_text_range: array (nullable = true)
|    |    |-- element: long (containsNull = true)
|    |-- entities: struct (nullable = true)
|    |    |-- hashtags: array (nullable = true)
|    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |-- text: string (nullable = true)
|    |    |-- media: array (nullable = true)
|    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |-- additional_media_info: struct (nullable = true)
|    |    |    |    |    |-- monetizable: boolean (nullable = true)
|    |    |    |    |-- display_url: string (nullable = true)
|    |    |    |    |-- expanded_url: string (nullable = true)
|    |    |    |    |-- id: long (nullable = true)
|    |    |    |    |-- id_str: string (nullable = true)
|    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |-- media_url: string (nullable = true)
|    |    |    |    |-- media_url_https: string (nullable = true)
|    |    |    |    |-- sizes: struct (nullable = true)
|    |    |    |    |    |-- large: struct (nullable = true)
|    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |-- medium: struct (nullable = true)
```

```
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |-- small: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |-- thumb: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |-- source_status_id: long (nullable = true)
 |    |    |    |    |-- source_status_id_str: string (nullable = true)
 |    |    |    |    |-- source_user_id: long (nullable = true)
 |    |    |    |    |-- source_user_id_str: string (nullable = true)
 |    |    |    |    |-- type: string (nullable = true)
 |    |    |    |    |-- url: string (nullable = true)
 |    |    |-- symbols: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- text: string (nullable = true)
 |    |    |-- urls: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- url: string (nullable = true)
 |    |    |-- user_mentions: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- id: long (nullable = true)
 |    |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- name: string (nullable = true)
 |    |    |    |    |-- screen_name: string (nullable = true)
 |    |-- extended_entities: struct (nullable = true)
 |    |    |-- media: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |    |-- id: long (nullable = true)
 |    |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- media_url: string (nullable = true)
 |    |    |    |    |-- media_url_https: string (nullable = true)
 |    |    |    |    |-- sizes: struct (nullable = true)
 |    |    |    |    |    |-- large: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |-- medium: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |-- small: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |-- thumb: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |-- source_status_id: long (nullable = true)
 |    |    |    |    |-- source_status_id_str: string (nullable = true)
 |    |    |    |    |-- source_user_id: long (nullable = true)
 |    |    |    |    |-- source_user_id_str: string (nullable = true)
 |    |    |    |    |-- type: string (nullable = true)
 |    |    |    |    |-- url: string (nullable = true)
 |    |    |    |    |-- video_info: struct (nullable = true)
 |    |    |    |    |    |-- aspect_ratio: array (nullable = true)
 |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |-- duration_millis: long (nullable = true)
 |    |    |    |    |    |-- variants: array (nullable = true)
 |    |    |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |    |    |-- bitrate: long (nullable = true)
 |    |    |    |    |    |    |    |-- content_type: string (nullable = true)
 |    |    |    |    |    |    |    |-- url: string (nullable = true)
 |    |-- extended_tweet: struct (nullable = true)
 |    |    |-- display_text_range: array (nullable = true)
 |    |    |    |-- element: long (containsNull = true)
 |    |    |-- entities: struct (nullable = true)
 |    |    |    |-- hashtags: array (nullable = true)
 |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |-- text: string (nullable = true)
 |    |    |    |-- media: array (nullable = true)
 |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |    |    |-- description: string (nullable = true)
 |    |    |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |    |    |-- id: long (nullable = true)
 |    |    |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |-- media_url: string (nullable = true)
 |    |    |    |    |    |-- media_url_https: string (nullable = true)
 |    |    |    |    |    |-- sizes: struct (nullable = true)
 |    |    |    |    |    |    |-- large: struct (nullable = true)
 |    |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |    |-- medium: struct (nullable = true)
 |    |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |    |-- small: struct (nullable = true)
 |    |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |    |-- thumb: struct (nullable = true)
 |    |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |-- source_status_id: long (nullable = true)
 |    |    |    |    |    |-- source_status_id_str: string (nullable = true)
 |    |    |    |    |    |-- source_user_id: long (nullable = true)
 |    |    |    |    |    |-- source_user_id_str: string (nullable = true)
 |    |    |    |    |    |-- type: string (nullable = true)
 |    |    |    |    |    |-- url: string (nullable = true)
 |    |    |    |    |    |-- video_info: struct (nullable = true)
 |    |    |    |    |    |    |-- aspect_ratio: array (nullable = true)
 |    |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |    |-- duration_millis: long (nullable = true)
 |    |    |    |    |    |    |-- variants: array (nullable = true)
 |    |    |    |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |    |    |    |-- bitrate: long (nullable = true)
 |    |    |    |    |    |    |    |    |-- content_type: string (nullable = true)
 |    |    |    |    |    |    |    |    |-- url: string (nullable = true)
 |    |    |    |-- symbols: array (nullable = true)
 |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |-- text: string (nullable = true)
 |    |    |    |-- urls: array (nullable = true)
 |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |-- url: string (nullable = true)
 |    |    |    |-- user_mentions: array (nullable = true)
 |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |-- id: long (nullable = true)
 |    |    |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |-- name: string (nullable = true)
 |    |    |    |    |    |-- screen_name: string (nullable = true)
 |    |    |-- extended_entities: struct (nullable = true)
 |    |    |    |-- media: array (nullable = true)
 |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |    |    |-- description: string (nullable = true)
 |    |    |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |    |    |-- id: long (nullable = true)
 |    |    |    |    |    |-- id_str: string (nullable = true)
```

```
|    |    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |-- media_url: string (nullable = true)
|    |    |    |    |    |-- media_url_https: string (nullable = true)
|    |    |    |    |    |-- sizes: struct (nullable = true)
|    |    |    |    |    |    |-- large: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- medium: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- small: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- thumb: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |-- source_status_id: long (nullable = true)
|    |    |    |    |    |-- source_status_id_str: string (nullable = true)
|    |    |    |    |    |-- source_user_id: long (nullable = true)
|    |    |    |    |    |-- source_user_id_str: string (nullable = true)
|    |    |    |    |    |-- type: string (nullable = true)
|    |    |    |    |    |-- url: string (nullable = true)
|    |    |    |    |    |-- video_info: struct (nullable = true)
|    |    |    |    |    |    |-- aspect_ratio: array (nullable = true)
|    |    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |    |-- duration_millis: long (nullable = true)
|    |    |    |    |    |    |-- variants: array (nullable = true)
|    |    |    |    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |    |    |    |-- bitrate: long (nullable = true)
|    |    |    |    |    |    |    |    |-- content_type: string (nullable = true)
|    |    |    |    |    |    |    |    |-- url: string (nullable = true)
|    |    |-- full_text: string (nullable = true)
|    |-- favorite_count: long (nullable = true)
|    |-- favorited: boolean (nullable = true)
|    |-- filter_level: string (nullable = true)
|    |-- id: long (nullable = true)
|    |-- id_str: string (nullable = true)
|    |-- in_reply_to_screen_name: string (nullable = true)
|    |-- in_reply_to_status_id: long (nullable = true)
|    |-- in_reply_to_status_id_str: string (nullable = true)
|    |-- in_reply_to_user_id: long (nullable = true)
|    |-- in_reply_to_user_id_str: string (nullable = true)
|    |-- is_quote_status: boolean (nullable = true)
|    |-- lang: string (nullable = true)
|    |-- place: struct (nullable = true)
|    |    |-- bounding_box: struct (nullable = true)
|    |    |    |-- coordinates: array (nullable = true)
|    |    |    |    |-- element: array (containsNull = true)
|    |    |    |    |    |-- element: array (containsNull = true)
|    |    |    |    |    |    |-- element: double (containsNull = true)
|    |    |    |-- type: string (nullable = true)
|    |    |-- country: string (nullable = true)
|    |    |-- country_code: string (nullable = true)
|    |    |-- full_name: string (nullable = true)
|    |    |-- id: string (nullable = true)
|    |    |-- name: string (nullable = true)
|    |    |-- place_type: string (nullable = true)
|    |    |-- url: string (nullable = true)
|    |-- possibly_sensitive: boolean (nullable = true)
|    |-- quote_count: long (nullable = true)
|    |-- quoted_status_id: long (nullable = true)
|    |-- quoted_status_id_str: string (nullable = true)
|    |-- reply_count: long (nullable = true)
|    |-- retweet_count: long (nullable = true)
|    |-- retweeted: boolean (nullable = true)
|    |-- source: string (nullable = true)
|    |-- text: string (nullable = true)
|    |-- truncated: boolean (nullable = true)
|    |-- user: struct (nullable = true)
|    |    |-- contributors_enabled: boolean (nullable = true)
|    |    |-- created_at: string (nullable = true)
|    |    |-- default_profile: boolean (nullable = true)
|    |    |-- default_profile_image: boolean (nullable = true)
|    |    |-- description: string (nullable = true)
|    |    |-- favourites_count: long (nullable = true)
|    |    |-- followers_count: long (nullable = true)
|    |    |-- friends_count: long (nullable = true)
|    |    |-- geo_enabled: boolean (nullable = true)
|    |    |-- id: long (nullable = true)
|    |    |-- id_str: string (nullable = true)
|    |    |-- is_translator: boolean (nullable = true)
|    |    |-- listed_count: long (nullable = true)
|    |    |-- location: string (nullable = true)
|    |    |-- name: string (nullable = true)
|    |    |-- profile_background_color: string (nullable = true)
|    |    |-- profile_background_image_url: string (nullable = true)
|    |    |-- profile_background_image_url_https: string (nullable = true)
|    |    |-- profile_background_tile: boolean (nullable = true)
|    |    |-- profile_banner_url: string (nullable = true)
|    |    |-- profile_image_url: string (nullable = true)
|    |    |-- profile_image_url_https: string (nullable = true)
|    |    |-- profile_link_color: string (nullable = true)
|    |    |-- profile_sidebar_border_color: string (nullable = true)
|    |    |-- profile_sidebar_fill_color: string (nullable = true)
|    |    |-- profile_text_color: string (nullable = true)
|    |    |-- profile_use_background_image: boolean (nullable = true)
|    |    |-- protected: boolean (nullable = true)
|    |    |-- screen_name: string (nullable = true)
|    |    |-- statuses_count: long (nullable = true)
|    |    |-- translator_type: string (nullable = true)
|    |    |-- url: string (nullable = true)
|    |    |-- verified: boolean (nullable = true)
|    |    |-- withheld_in_countries: array (nullable = true)
|    |    |    |-- element: string (containsNull = true)
|-- quoted_status_id: long (nullable = true)
|-- quoted_status_id_str: string (nullable = true)
|-- quoted_status_permalink: struct (nullable = true)
|    |-- display: string (nullable = true)
|    |-- expanded: string (nullable = true)
|    |-- url: string (nullable = true)
|-- quoted_text: string (nullable = true)
|-- reply_count: long (nullable = true)
|-- retweet_count: long (nullable = true)
|-- retweeted: string (nullable = true)
|-- retweeted_from: string (nullable = true)
|-- retweeted_status: struct (nullable = true)
|    |-- coordinates: struct (nullable = true)
|    |    |-- coordinates: array (nullable = true)
|    |    |    |-- element: double (containsNull = true)
|    |    |-- type: string (nullable = true)
|    |-- created_at: string (nullable = true)
|    |-- display_text_range: array (nullable = true)
|    |    |-- element: long (containsNull = true)
|    |-- entities: struct (nullable = true)
|    |    |-- hashtags: array (nullable = true)
|    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |-- text: string (nullable = true)
|    |    |-- media: array (nullable = true)
|    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |-- additional_media_info: struct (nullable = true)
|    |    |    |    |    |-- monetizable: boolean (nullable = true)
|    |    |    |    |-- display_url: string (nullable = true)
|    |    |    |    |-- expanded_url: string (nullable = true)
|    |    |    |    |-- id: long (nullable = true)
|    |    |    |    |-- id_str: string (nullable = true)
|    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |-- media_url: string (nullable = true)
|    |    |    |    |-- media_url_https: string (nullable = true)
|    |    |    |    |-- sizes: struct (nullable = true)
|    |    |    |    |    |-- large: struct (nullable = true)
|    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |-- medium: struct (nullable = true)
|    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |-- small: struct (nullable = true)
|    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |-- thumb: struct (nullable = true)
|    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |-- source_status_id: long (nullable = true)
|    |    |    |    |-- source_status_id_str: string (nullable = true)
|    |    |    |    |-- source_user_id: long (nullable = true)
```

```
 |    |    |    |    |-- source_user_id_str: string (nullable = true)
 |    |    |    |    |-- type: string (nullable = true)
 |    |    |    |    |-- url: string (nullable = true)
 |    |    |-- symbols: array (nullable = true)
 |    |    |    |-- element: string (containsNull = true)
 |    |    |-- urls: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- url: string (nullable = true)
 |    |    |-- user_mentions: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- id: long (nullable = true)
 |    |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- name: string (nullable = true)
 |    |    |    |    |-- screen_name: string (nullable = true)
 |    |-- extended_entities: struct (nullable = true)
 |    |    |-- media: array (nullable = true)
 |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |    |-- id: long (nullable = true)
 |    |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |-- media_url: string (nullable = true)
 |    |    |    |    |-- media_url_https: string (nullable = true)
 |    |    |    |    |-- sizes: struct (nullable = true)
 |    |    |    |    |    |-- large: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |-- medium: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |-- small: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |-- thumb: struct (nullable = true)
 |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |-- source_status_id: long (nullable = true)
 |    |    |    |    |-- source_status_id_str: string (nullable = true)
 |    |    |    |    |-- source_user_id: long (nullable = true)
 |    |    |    |    |-- source_user_id_str: string (nullable = true)
 |    |    |    |    |-- type: string (nullable = true)
 |    |    |    |    |-- url: string (nullable = true)
 |    |    |    |    |-- video_info: struct (nullable = true)
 |    |    |    |    |    |-- aspect_ratio: array (nullable = true)
 |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |-- duration_millis: long (nullable = true)
 |    |    |    |    |    |-- variants: array (nullable = true)
 |    |    |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |    |    |-- bitrate: long (nullable = true)
 |    |    |    |    |    |    |    |-- content_type: string (nullable = true)
 |    |    |    |    |    |    |    |-- url: string (nullable = true)
 |    |-- extended_tweet: struct (nullable = true)
 |    |    |-- display_text_range: array (nullable = true)
 |    |    |    |-- element: long (containsNull = true)
 |    |    |-- entities: struct (nullable = true)
 |    |    |    |-- hashtags: array (nullable = true)
 |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |-- text: string (nullable = true)
 |    |    |    |-- media: array (nullable = true)
 |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |    |    |    |-- description: string (nullable = true)
 |    |    |    |    |    |    |-- embeddable: boolean (nullable = true)
 |    |    |    |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |    |    |    |-- title: string (nullable = true)
 |    |    |    |    |    |-- description: string (nullable = true)
 |    |    |    |    |    |-- display_url: string (nullable = true)
```

```
|    |    |    |    |    |-- expanded_url: string (nullable = true)
|    |    |    |    |    |-- id: long (nullable = true)
|    |    |    |    |    |-- id_str: string (nullable = true)
|    |    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |-- media_url: string (nullable = true)
|    |    |    |    |    |-- media_url_https: string (nullable = true)
|    |    |    |    |    |-- sizes: struct (nullable = true)
|    |    |    |    |    |    |-- large: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- medium: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- small: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- thumb: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |-- source_status_id: long (nullable = true)
|    |    |    |    |    |-- source_status_id_str: string (nullable = true)
|    |    |    |    |    |-- source_user_id: long (nullable = true)
|    |    |    |    |    |-- source_user_id_str: string (nullable = true)
|    |    |    |    |    |-- type: string (nullable = true)
|    |    |    |    |    |-- url: string (nullable = true)
|    |    |    |    |    |-- video_info: struct (nullable = true)
|    |    |    |    |    |    |-- aspect_ratio: array (nullable = true)
|    |    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |    |-- duration_millis: long (nullable = true)
|    |    |    |    |    |    |-- variants: array (nullable = true)
|    |    |    |    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |    |    |    |-- bitrate: long (nullable = true)
|    |    |    |    |    |    |    |    |-- content_type: string (nullable = true)
|    |    |    |    |    |    |    |    |-- url: string (nullable = true)
|    |    |    |-- symbols: array (nullable = true)
|    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |-- text: string (nullable = true)
|    |    |    |-- urls: array (nullable = true)
|    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |-- display_url: string (nullable = true)
|    |    |    |    |    |-- expanded_url: string (nullable = true)
|    |    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |-- url: string (nullable = true)
|    |    |    |-- user_mentions: array (nullable = true)
|    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |-- id: long (nullable = true)
|    |    |    |    |    |-- id_str: string (nullable = true)
|    |    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |-- name: string (nullable = true)
|    |    |    |    |    |-- screen_name: string (nullable = true)
|    |    |-- extended_entities: struct (nullable = true)
|    |    |    |-- media: array (nullable = true)
|    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |-- additional_media_info: struct (nullable = true)
|    |    |    |    |    |    |-- description: string (nullable = true)
|    |    |    |    |    |    |-- embeddable: boolean (nullable = true)
|    |    |    |    |    |    |-- monetizable: boolean (nullable = true)
|    |    |    |    |    |    |-- title: string (nullable = true)
|    |    |    |    |    |-- description: string (nullable = true)
|    |    |    |    |    |-- display_url: string (nullable = true)
|    |    |    |    |    |-- expanded_url: string (nullable = true)
|    |    |    |    |    |-- id: long (nullable = true)
|    |    |    |    |    |-- id_str: string (nullable = true)
|    |    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |-- media_url: string (nullable = true)
|    |    |    |    |    |-- media_url_https: string (nullable = true)
|    |    |    |    |    |-- sizes: struct (nullable = true)
|    |    |    |    |    |    |-- large: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- medium: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- small: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- thumb: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |-- source_status_id: long (nullable = true)
|    |    |    |    |    |-- source_status_id_str: string (nullable = true)
|    |    |    |    |    |-- source_user_id: long (nullable = true)
|    |    |    |    |    |-- source_user_id_str: string (nullable = true)
|    |    |    |    |    |-- type: string (nullable = true)
|    |    |    |    |    |-- url: string (nullable = true)
|    |    |    |    |    |-- video_info: struct (nullable = true)
|    |    |    |    |    |    |-- aspect_ratio: array (nullable = true)
|    |    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |    |-- duration_millis: long (nullable = true)
|    |    |    |    |    |    |-- variants: array (nullable = true)
|    |    |    |    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |    |    |    |-- bitrate: long (nullable = true)
|    |    |    |    |    |    |    |    |-- content_type: string (nullable = true)
|    |    |    |    |    |    |    |    |-- url: string (nullable = true)
|    |    |-- full_text: string (nullable = true)
|    |-- favorite_count: long (nullable = true)
|    |-- favorited: boolean (nullable = true)
|    |-- filter_level: string (nullable = true)
|    |-- geo: struct (nullable = true)
|    |    |-- coordinates: array (nullable = true)
|    |    |    |-- element: double (containsNull = true)
|    |    |-- type: string (nullable = true)
|    |-- id: long (nullable = true)
|    |-- id_str: string (nullable = true)
|    |-- in_reply_to_screen_name: string (nullable = true)
|    |-- in_reply_to_status_id: long (nullable = true)
|    |-- in_reply_to_status_id_str: string (nullable = true)
|    |-- in_reply_to_user_id: long (nullable = true)
|    |-- in_reply_to_user_id_str: string (nullable = true)
|    |-- is_quote_status: boolean (nullable = true)
|    |-- lang: string (nullable = true)
|    |-- place: struct (nullable = true)
|    |    |-- bounding_box: struct (nullable = true)
|    |    |    |-- coordinates: array (nullable = true)
|    |    |    |    |-- element: array (containsNull = true)
|    |    |    |    |    |-- element: array (containsNull = true)
|    |    |    |    |    |    |-- element: double (containsNull = true)
|    |    |    |-- type: string (nullable = true)
|    |    |-- country: string (nullable = true)
|    |    |-- country_code: string (nullable = true)
|    |    |-- full_name: string (nullable = true)
|    |    |-- id: string (nullable = true)
|    |    |-- name: string (nullable = true)
|    |    |-- place_type: string (nullable = true)
|    |    |-- url: string (nullable = true)
|    |-- possibly_sensitive: boolean (nullable = true)
|    |-- quote_count: long (nullable = true)
|    |-- quoted_status: struct (nullable = true)
|    |    |-- created_at: string (nullable = true)
|    |    |-- display_text_range: array (nullable = true)
|    |    |    |-- element: long (containsNull = true)
|    |    |-- entities: struct (nullable = true)
|    |    |    |-- hashtags: array (nullable = true)
|    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |-- text: string (nullable = true)
|    |    |    |-- media: array (nullable = true)
|    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |-- additional_media_info: struct (nullable = true)
|    |    |    |    |    |    |-- monetizable: boolean (nullable = true)
|    |    |    |    |    |-- display_url: string (nullable = true)
|    |    |    |    |    |-- expanded_url: string (nullable = true)
|    |    |    |    |    |-- id: long (nullable = true)
|    |    |    |    |    |-- id_str: string (nullable = true)
|    |    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |-- media_url: string (nullable = true)
|    |    |    |    |    |-- media_url_https: string (nullable = true)
|    |    |    |    |    |-- sizes: struct (nullable = true)
|    |    |    |    |    |    |-- large: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- medium: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- small: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- thumb: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |-- source_status_id: long (nullable = true)
|    |    |    |    |    |-- source_status_id_str: string (nullable = true)
|    |    |    |    |    |-- source_user_id: long (nullable = true)
|    |    |    |    |    |-- source_user_id_str: string (nullable = true)
|    |    |    |    |    |-- type: string (nullable = true)
|    |    |    |    |    |-- url: string (nullable = true)
|    |    |    |-- symbols: array (nullable = true)
|    |    |    |    |-- element: string (containsNull = true)
|    |    |    |-- urls: array (nullable = true)
|    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |-- display_url: string (nullable = true)
|    |    |    |    |    |-- expanded_url: string (nullable = true)
|    |    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |-- url: string (nullable = true)
|    |    |    |-- user_mentions: array (nullable = true)
|    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |-- id: long (nullable = true)
|    |    |    |    |    |-- id_str: string (nullable = true)
|    |    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |-- name: string (nullable = true)
|    |    |    |    |    |-- screen_name: string (nullable = true)
|    |    |-- extended_entities: struct (nullable = true)
|    |    |    |-- media: array (nullable = true)
|    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |-- additional_media_info: struct (nullable = true)
|    |    |    |    |    |    |-- monetizable: boolean (nullable = true)
|    |    |    |    |    |-- display_url: string (nullable = true)
|    |    |    |    |    |-- expanded_url: string (nullable = true)
|    |    |    |    |    |-- id: long (nullable = true)
|    |    |    |    |    |-- id_str: string (nullable = true)
|    |    |    |    |    |-- indices: array (nullable = true)
|    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |-- media_url: string (nullable = true)
|    |    |    |    |    |-- media_url_https: string (nullable = true)
|    |    |    |    |    |-- sizes: struct (nullable = true)
|    |    |    |    |    |    |-- large: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- medium: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- small: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |    |-- thumb: struct (nullable = true)
|    |    |    |    |    |    |    |-- h: long (nullable = true)
|    |    |    |    |    |    |    |-- resize: string (nullable = true)
|    |    |    |    |    |    |    |-- w: long (nullable = true)
|    |    |    |    |    |-- source_status_id: long (nullable = true)
|    |    |    |    |    |-- source_status_id_str: string (nullable = true)
|    |    |    |    |    |-- source_user_id: long (nullable = true)
|    |    |    |    |    |-- source_user_id_str: string (nullable = true)
|    |    |    |    |    |-- type: string (nullable = true)
|    |    |    |    |    |-- url: string (nullable = true)
|    |    |    |    |    |-- video_info: struct (nullable = true)
|    |    |    |    |    |    |-- aspect_ratio: array (nullable = true)
|    |    |    |    |    |    |    |-- element: long (containsNull = true)
|    |    |    |    |    |    |-- duration_millis: long (nullable = true)
|    |    |    |    |    |    |-- variants: array (nullable = true)
|    |    |    |    |    |    |    |-- element: struct (containsNull = true)
|    |    |    |    |    |    |    |    |-- bitrate: long (nullable = true)
|    |    |    |    |    |    |    |    |-- content_type: string (nullable = true)
|    |    |    |    |    |    |    |    |-- url: string (nullable = true)
|    |    |-- extended_tweet: struct (nullable = true)
```

```
 |    |    |    |-- display_text_range: array (nullable = true)
 |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |-- entities: struct (nullable = true)
 |    |    |    |    |-- hashtags: array (nullable = true)
 |    |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |    |-- text: string (nullable = true)
 |    |    |    |    |-- media: array (nullable = true)
 |    |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |    |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |    |    |    |-- description: string (nullable = true)
 |    |    |    |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |    |    |    |-- id: long (nullable = true)
 |    |    |    |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |    |-- media_url: string (nullable = true)
 |    |    |    |    |    |    |-- media_url_https: string (nullable = true)
 |    |    |    |    |    |    |-- sizes: struct (nullable = true)
 |    |    |    |    |    |    |    |-- large: struct (nullable = true)
 |    |    |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |    |    |-- medium: struct (nullable = true)
 |    |    |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |    |    |-- small: struct (nullable = true)
 |    |    |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |    |    |-- thumb: struct (nullable = true)
 |    |    |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |    |-- type: string (nullable = true)
 |    |    |    |    |    |    |-- url: string (nullable = true)
 |    |    |    |    |    |    |-- video_info: struct (nullable = true)
 |    |    |    |    |    |    |    |-- aspect_ratio: array (nullable = true)
 |    |    |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |    |    |-- duration_millis: long (nullable = true)
 |    |    |    |    |    |    |    |-- variants: array (nullable = true)
 |    |    |    |    |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |    |    |    |    |-- bitrate: long (nullable = true)
 |    |    |    |    |    |    |    |    |    |-- content_type: string (nullable = true)
 |    |    |    |    |    |    |    |    |    |-- url: string (nullable = true)
 |    |    |    |    |-- symbols: array (nullable = true)
 |    |    |    |    |    |-- element: string (containsNull = true)
 |    |    |    |    |-- urls: array (nullable = true)
 |    |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |    |-- url: string (nullable = true)
 |    |    |    |    |-- user_mentions: array (nullable = true)
 |    |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |    |-- id: long (nullable = true)
 |    |    |    |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |    |-- name: string (nullable = true)
 |    |    |    |    |    |    |-- screen_name: string (nullable = true)
 |    |    |    |-- extended_entities: struct (nullable = true)
 |    |    |    |    |-- media: array (nullable = true)
 |    |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |    |-- additional_media_info: struct (nullable = true)
 |    |    |    |    |    |    |    |-- monetizable: boolean (nullable = true)
 |    |    |    |    |    |    |-- description: string (nullable = true)
 |    |    |    |    |    |    |-- display_url: string (nullable = true)
 |    |    |    |    |    |    |-- expanded_url: string (nullable = true)
 |    |    |    |    |    |    |-- id: long (nullable = true)
 |    |    |    |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |    |    |    |-- indices: array (nullable = true)
 |    |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |    |-- media_url: string (nullable = true)
 |    |    |    |    |    |    |-- media_url_https: string (nullable = true)
 |    |    |    |    |    |    |-- sizes: struct (nullable = true)
 |    |    |    |    |    |    |    |-- large: struct (nullable = true)
 |    |    |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |    |    |-- medium: struct (nullable = true)
 |    |    |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |    |    |-- small: struct (nullable = true)
 |    |    |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |    |    |-- thumb: struct (nullable = true)
 |    |    |    |    |    |    |    |    |-- h: long (nullable = true)
 |    |    |    |    |    |    |    |    |-- resize: string (nullable = true)
 |    |    |    |    |    |    |    |    |-- w: long (nullable = true)
 |    |    |    |    |    |    |-- type: string (nullable = true)
 |    |    |    |    |    |    |-- url: string (nullable = true)
 |    |    |    |    |    |    |-- video_info: struct (nullable = true)
 |    |    |    |    |    |    |    |-- aspect_ratio: array (nullable = true)
 |    |    |    |    |    |    |    |    |-- element: long (containsNull = true)
 |    |    |    |    |    |    |    |-- duration_millis: long (nullable = true)
 |    |    |    |    |    |    |    |-- variants: array (nullable = true)
 |    |    |    |    |    |    |    |    |-- element: struct (containsNull = true)
 |    |    |    |    |    |    |    |    |    |-- bitrate: long (nullable = true)
 |    |    |    |    |    |    |    |    |    |-- content_type: string (nullable = true)
 |    |    |    |    |    |    |    |    |    |-- url: string (nullable = true)
 |    |    |    |-- full_text: string (nullable = true)
 |    |    |-- favorite_count: long (nullable = true)
 |    |    |-- favorited: boolean (nullable = true)
 |    |    |-- filter_level: string (nullable = true)
 |    |    |-- id: long (nullable = true)
 |    |    |-- id_str: string (nullable = true)
 |    |    |-- in_reply_to_screen_name: string (nullable = true)
 |    |    |-- in_reply_to_status_id: long (nullable = true)
 |    |    |-- in_reply_to_status_id_str: string (nullable = true)
 |    |    |-- in_reply_to_user_id: long (nullable = true)
 |    |    |-- in_reply_to_user_id_str: string (nullable = true)
 |    |    |-- is_quote_status: boolean (nullable = true)
 |    |    |-- lang: string (nullable = true)
 |    |    |-- place: struct (nullable = true)
 |    |    |    |-- bounding_box: struct (nullable = true)
 |    |    |    |    |-- coordinates: array (nullable = true)
 |    |    |    |    |    |-- element: array (containsNull = true)
 |    |    |    |    |    |    |-- element: array (containsNull = true)
 |    |    |    |    |    |    |    |-- element: double (containsNull = true)
 |    |    |    |    |-- type: string (nullable = true)
 |    |    |    |-- country: string (nullable = true)
 |    |    |    |-- country_code: string (nullable = true)
 |    |    |    |-- full_name: string (nullable = true)
 |    |    |    |-- id: string (nullable = true)
 |    |    |    |-- name: string (nullable = true)
 |    |    |    |-- place_type: string (nullable = true)
 |    |    |    |-- url: string (nullable = true)
 |    |    |-- possibly_sensitive: boolean (nullable = true)
 |    |    |-- quote_count: long (nullable = true)
 |    |    |-- quoted_status_id: long (nullable = true)
 |    |    |-- quoted_status_id_str: string (nullable = true)
 |    |    |-- reply_count: long (nullable = true)
 |    |    |-- retweet_count: long (nullable = true)
 |    |    |-- retweeted: boolean (nullable = true)
 |    |    |-- source: string (nullable = true)
 |    |    |-- text: string (nullable = true)
 |    |    |-- truncated: boolean (nullable = true)
 |    |    |-- user: struct (nullable = true)
 |    |    |    |-- contributors_enabled: boolean (nullable = true)
 |    |    |    |-- created_at: string (nullable = true)
 |    |    |    |-- default_profile: boolean (nullable = true)
 |    |    |    |-- default_profile_image: boolean (nullable = true)
 |    |    |    |-- description: string (nullable = true)
 |    |    |    |-- favourites_count: long (nullable = true)
 |    |    |    |-- followers_count: long (nullable = true)
 |    |    |    |-- friends_count: long (nullable = true)
 |    |    |    |-- geo_enabled: boolean (nullable = true)
 |    |    |    |-- id: long (nullable = true)
 |    |    |    |-- id_str: string (nullable = true)
 |    |    |    |-- is_translator: boolean (nullable = true)
 |    |    |    |-- listed_count: long (nullable = true)
 |    |    |    |-- location: string (nullable = true)
 |    |    |    |-- name: string (nullable = true)
 |    |    |    |-- profile_background_color: string (nullable = true)
 |    |    |    |-- profile_background_image_url: string (nullable = true)
 |    |    |    |-- profile_background_image_url_https: string (nullable = true)
 |    |    |    |-- profile_background_tile: boolean (nullable = true)
 |    |    |    |-- profile_banner_url: string (nullable = true)
```

```
|    |    |    |-- profile_image_url: string (nullable = true)
|    |    |    |-- profile_image_url_https: string (nullable = true)
|    |    |    |-- profile_link_color: string (nullable = true)
|    |    |    |-- profile_sidebar_border_color: string (nullable = true)
|    |    |    |-- profile_sidebar_fill_color: string (nullable = true)
|    |    |    |-- profile_text_color: string (nullable = true)
|    |    |    |-- profile_use_background_image: boolean (nullable = true)
|    |    |    |-- protected: boolean (nullable = true)
|    |    |    |-- screen_name: string (nullable = true)
|    |    |    |-- statuses_count: long (nullable = true)
|    |    |    |-- translator_type: string (nullable = true)
|    |    |    |-- url: string (nullable = true)
|    |    |    |-- verified: boolean (nullable = true)
|    |    |    |-- withheld_in_countries: array (nullable = true)
|    |    |    |    |-- element: string (containsNull = true)
|    |-- quoted_status_id: long (nullable = true)
|    |-- quoted_status_id_str: string (nullable = true)
|    |-- quoted_status_permalink: struct (nullable = true)
|    |    |-- display: string (nullable = true)
|    |    |-- expanded: string (nullable = true)
|    |    |-- url: string (nullable = true)
|    |-- reply_count: long (nullable = true)
|    |-- retweet_count: long (nullable = true)
|    |-- retweeted: boolean (nullable = true)
|    |-- scopes: struct (nullable = true)
|    |    |-- followers: boolean (nullable = true)
|    |-- source: string (nullable = true)
|    |-- text: string (nullable = true)
|    |-- truncated: boolean (nullable = true)
|    |-- user: struct (nullable = true)
|    |    |-- contributors_enabled: boolean (nullable = true)
|    |    |-- created_at: string (nullable = true)
|    |    |-- default_profile: boolean (nullable = true)
|    |    |-- default_profile_image: boolean (nullable = true)
|    |    |-- description: string (nullable = true)
|    |    |-- favourites_count: long (nullable = true)
|    |    |-- followers_count: long (nullable = true)
|    |    |-- friends_count: long (nullable = true)
|    |    |-- geo_enabled: boolean (nullable = true)
|    |    |-- id: long (nullable = true)
|    |    |-- id_str: string (nullable = true)
|    |    |-- is_translator: boolean (nullable = true)
|    |    |-- listed_count: long (nullable = true)
|    |    |-- location: string (nullable = true)
|    |    |-- name: string (nullable = true)
|    |    |-- profile_background_color: string (nullable = true)
|    |    |-- profile_background_image_url: string (nullable = true)
|    |    |-- profile_background_image_url_https: string (nullable = true)
|    |    |-- profile_background_tile: boolean (nullable = true)
|    |    |-- profile_banner_url: string (nullable = true)
|    |    |-- profile_image_url: string (nullable = true)
|    |    |-- profile_image_url_https: string (nullable = true)
|    |    |-- profile_link_color: string (nullable = true)
|    |    |-- profile_sidebar_border_color: string (nullable = true)
|    |    |-- profile_sidebar_fill_color: string (nullable = true)
|    |    |-- profile_text_color: string (nullable = true)
|    |    |-- profile_use_background_image: boolean (nullable = true)
|    |    |-- protected: boolean (nullable = true)
|    |    |-- screen_name: string (nullable = true)
|    |    |-- statuses_count: long (nullable = true)
|    |    |-- translator_type: string (nullable = true)
|    |    |-- url: string (nullable = true)
|    |    |-- verified: boolean (nullable = true)
|    |    |-- withheld_in_countries: array (nullable = true)
|    |    |    |-- element: string (containsNull = true)
|-- source: string (nullable = true)
|-- text: string (nullable = true)
|-- timestamp_ms: string (nullable = true)
|-- truncated: boolean (nullable = true)
|-- tweet_text: string (nullable = true)
|-- user: struct (nullable = true)
|    |-- contributors_enabled: boolean (nullable = true)
|    |-- created_at: string (nullable = true)
|    |-- default_profile: boolean (nullable = true)
|    |-- default_profile_image: boolean (nullable = true)
|    |-- description: string (nullable = true)
|    |-- favourites_count: long (nullable = true)
|    |-- followers_count: long (nullable = true)
|    |-- friends_count: long (nullable = true)
|    |-- geo_enabled: boolean (nullable = true)
|    |-- id: long (nullable = true)
|    |-- id_str: string (nullable = true)
|    |-- is_translator: boolean (nullable = true)
```

```
 |    |-- listed_count: long (nullable = true)
 |    |-- location: string (nullable = true)
 |    |-- name: string (nullable = true)
 |    |-- profile_background_color: string (nullable = true)
 |    |-- profile_background_image_url: string (nullable = true)
 |    |-- profile_background_image_url_https: string (nullable = true)
 |    |-- profile_background_tile: boolean (nullable = true)
 |    |-- profile_banner_url: string (nullable = true)
 |    |-- profile_image_url: string (nullable = true)
 |    |-- profile_image_url_https: string (nullable = true)
 |    |-- profile_link_color: string (nullable = true)
 |    |-- profile_sidebar_border_color: string (nullable = true)
 |    |-- profile_sidebar_fill_color: string (nullable = true)
 |    |-- profile_text_color: string (nullable = true)
 |    |-- profile_use_background_image: boolean (nullable = true)
 |    |-- protected: boolean (nullable = true)
 |    |-- screen_name: string (nullable = true)
 |    |-- statuses_count: long (nullable = true)
 |    |-- translator_type: string (nullable = true)
 |    |-- url: string (nullable = true)
 |    |-- verified: boolean (nullable = true)
 |    |-- withheld_in_countries: array (nullable = true)
 |    |    |-- element: string (containsNull = true)
```

Whole Data

In [2]:

```
data = spark.read.json('gs://msca-bdp-tweets/final_project/')
```

```
22/12/04 03:42:06 WARN org.apache.spark.sql.execution.datasources.SharedInMemoryCache: Evicting cached table partition metadata from memory due to size constraints (spark.sql.hive.filesourcePartitionFileCacheSize = 262144000 bytes). This may impact query planning performance.
22/12/04 03:47:36 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'.
```

In [18]:

```
data.select(data.retweet_count).filter(data.retweet_count != 0).show()
```

```
[Stage 25:===================================================>(2267 + 9) / 2276]
+-------------+
|retweet_count|
+-------------+
+-------------+
```

In [19]:

```
data.select(col('retweeted_status').getItem('retweet_count')).show()
```

```
+------------------------------+
|retweeted_status.retweet_count|
+------------------------------+
|                          3008|
|                          null|
|                          2733|
|                          null|
|                            78|
|                          1856|
|                         23904|
|                             5|
|                           977|
|                          null|
|                             2|
|                          1765|
|                          2238|
|                            63|
|                          null|
|                             1|
|                           918|
|                             4|
|                          2239|
|                         13593|
+------------------------------+
only showing top 20 rows
```

In [19]:

```
data.select(col('retweeted_status').getItem('retweeted')).show()
```

```
+--------------------------+
|retweeted_status.retweeted|
+--------------------------+
|                     false|
|                      null|
|                     false|
|                      null|
|                     false|
|                     false|
|                     false|
|                     false|
|                     false|
|                      null|
|                     false|
|                     false|
|                     false|
|                     false|
|                      null|
|                     false|
|                     false|
|                     false|
|                     false|
|                     false|
+--------------------------+
only showing top 20 rows
```

In [81]:

```
data.select('retweet_count').orderBy(col('retweet_count').desc()).show()
```

```
22/12/01 06:12:05 WARN org.apache.spark.scheduler.cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Requesting
driver to remove executor 116 for reason Container marked as failed: container_1669695139548_0008_01_000118 on ho
st: hub-msca-bdp-dphub-students-chenfeng-sw-xf6k.c.msca-bdp-students.internal. Exit status: -100. Diagnostics: Co
ntainer released on a *lost* node.
22/12/01 06:12:05 ERROR org.apache.spark.scheduler.cluster.YarnScheduler: Lost executor 109 on hub-msca-bdp-dphub
-students-chenfeng-sw-xf6k.c.msca-bdp-students.internal: Container marked as failed: container_1669695139548_0008
_01_000111 on host: hub-msca-bdp-dphub-students-chenfeng-sw-xf6k.c.msca-bdp-students.internal. Exit status: -100.
Diagnostics: Container released on a *lost* node.
22/12/01 06:12:05 ERROR org.apache.spark.scheduler.cluster.YarnScheduler: Lost executor 116 on hub-msca-bdp-dphub
-students-chenfeng-sw-xf6k.c.msca-bdp-students.internal: Container marked as failed: container_1669695139548_0008
_01_000118 on host: hub-msca-bdp-dphub-students-chenfeng-sw-xf6k.c.msca-bdp-students.internal. Exit status: -100.
Diagnostics: Container released on a *lost* node.
22/12/01 06:12:05 WARN org.apache.spark.scheduler.cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Requesting
driver to remove executor 109 for reason Container marked as failed: container_1669695139548_0008_01_000111 on ho
st: hub-msca-bdp-dphub-students-chenfeng-sw-xf6k.c.msca-bdp-students.internal. Exit status: -100. Diagnostics: Co
ntainer released on a *lost* node.
[Stage 39:=====================================================>(5400 + 1) / 5401]
```

```
+-------------+
|retweet_count|
+-------------+
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
|            0|
+-------------+
only showing top 20 rows
```

In [73]:

```
data.select('retweeted','retweeted_from').show()
```

```
+---------+---------------+
|retweeted| retweeted_from|
+---------+---------------+
|       RT|            ABC|
|         |           null|
|       RT|     jaketapper|
|         |           null|
|       RT|      Josh_Moon|
|       RT|     Jim_Jordan|
|       RT|     meganbang3|
|       RT|   jewishaction|
|       RT|   LRiddickESPN|
|         |           null|
|       RT| Gisele23935327|
|       RT|  MichaelSteele|
|       RT|    ABCPolitics|
|       RT|     glowinasia|
|         |           null|
|       RT|  KaufmanAbrams|
|       RT|OccupyDemocrats|
|       RT|Meidas_Adrienne|
|       RT|    ABCPolitics|
|       RT|      KingJames|
+---------+---------------+
only showing top 20 rows
```

In [78]:

```
data.select('retweeted', 'retweeted_from','text',col('retweeted_status').getItem('retweet_count').alias('count'))
orderBy(col('count').desc()).show()
```

```
[Stage 37:====================================================>(5399 + 2) / 5401]
```

```
+---------+--------------+--------------------+------+
|retweeted|retweeted_from|                text| count|
+---------+--------------+--------------------+------+
|       RT|        nickjr|RT @nickjr: So ab...|516954|
|       RT|        nickjr|RT @nickjr: So ab...|516951|
|       RT|        nickjr|RT @nickjr: So ab...|516928|
|       RT|        nickjr|RT @nickjr: So ab...|516795|
|       RT|        nickjr|RT @nickjr: So ab...|516772|
|       RT|        nickjr|RT @nickjr: So ab...|516743|
|       RT|        nickjr|RT @nickjr: So ab...|516711|
|       RT|        nickjr|RT @nickjr: So ab...|516614|
|       RT|        nickjr|RT @nickjr: So ab...|516596|
|       RT|        nickjr|RT @nickjr: So ab...|516593|
|       RT|        nickjr|RT @nickjr: So ab...|516556|
|       RT|        nickjr|RT @nickjr: So ab...|516555|
|       RT|        nickjr|RT @nickjr: So ab...|516544|
|       RT|        nickjr|RT @nickjr: So ab...|516543|
|       RT|        nickjr|RT @nickjr: So ab...|516540|
|       RT|        nickjr|RT @nickjr: So ab...|516539|
|       RT|        nickjr|RT @nickjr: So ab...|516539|
|       RT|        nickjr|RT @nickjr: So ab...|516535|
|       RT|        nickjr|RT @nickjr: So ab...|516534|
|       RT|        nickjr|RT @nickjr: So ab...|516534|
+---------+--------------+--------------------+------+
only showing top 20 rows
```

In [50]:

```
data.select('user','place','created_at','retweet_count','coordinates','extended_tweet','text').limit(5).show()
```

```
[Stage 30:>                                                         (0 + 1) / 1]
```

```
+--------------------+-----+--------------------+-------------+-----------+--------------------+
|                user|place|          created_at|retweet_count|coordinates|      extended_tweet|
+--------------------+-----+--------------------+-------------+-----------+--------------------+
|{false, Thu Mar 0...| null|Tue May 24 22:09:...|            0|       null|                null|
|{false, Thu May 1...| null|Tue May 24 22:09:...|            0|       null|{[0, 219], {[], [...|
|{false, Fri Jun 2...| null|Tue May 24 22:09:...|            0|       null|                null|
|{false, Wed Feb 2...| null|Tue May 24 22:09:...|            0|       null|{[0, 182], {[{[0,...|
|{false, Wed Jul 0...| null|Tue May 24 22:09:...|            0|       null|                null|
+--------------------+-----+--------------------+-------------+-----------+--------------------+
```

In [3]:

```
new = data.select('user','created_at','extended_tweet','retweeted_status','text','id','retweeted_from')
```

In [4]:

```
new = new.withColumn("user_id", col("user").getItem("id")).\
        withColumn('user_name', col('user').getItem('screen_name')).\
        withColumn('user_loc',col('user').getItem('location')).\
        withColumn('user_descrip',col('user').getItem('description')).\
        withColumn('user_followerct',col('user').getItem('followers_count')).\
        select('user_id', 'user_name', 'user_loc', 'user_descrip','user_followerct',
               'created_at','text','extended_tweet','retweeted_status','id','retweeted_from')
```

In [27]:

```
# the place attributes is not useful since too many null
# new = new.filter(col("place").isNotNull()).\
#          withColumn("full_loc", col("place").getItem("full_name")).\
#          withColumn("country", col("place").getItem("country")).\
#          withColumn("location", col("place").getItem("name")).\
#          drop('place')
```

In [5]:

```
# Found the retweet count in the correct location
new = new.withColumn('retweet_ct', col('retweeted_status').getItem('retweet_count')).\
            withColumn('retweet', col('retweeted_status').getItem('retweeted')).\
            drop('retweeted_status')
```

In [6]:

```
# extract full text
new = new.withColumn('full_text',col('extended_tweet').getItem('full_text')).drop('extended_tweet')
```

In [94]:

```
new.printSchema()
```

```
root
 |-- user_id: long (nullable = true)
 |-- user_name: string (nullable = true)
 |-- user_loc: string (nullable = true)
 |-- user_favorct: long (nullable = true)
 |-- user_followerct: long (nullable = true)
 |-- created_at: string (nullable = true)
 |-- text: string (nullable = true)
 |-- id: long (nullable = true)
 |-- retweeted_from: string (nullable = true)
 |-- retweet_ct: long (nullable = true)
 |-- retweet: boolean (nullable = true)
 |-- full_text: string (nullable = true)
```

In [53]:

```
n = new.select('text','retweet_ct', 'retweet').filter(new.retweet_ct.isNotNull()).\
filter(col('retweet') == True).orderBy(col('retweet_ct').desc())
```

In [58]:

```
n = new.select('text','retweet_ct', 'retweet').filter(col('retweet') == True)
n.limit(10).collect()
```

Out[58]:

```
[Row(text='Indiana High School Softball | Live Streaming\nTBA vs Fishers\nEastbrook vs Oak Hill\nTBA vs Wabash\nS
outh Newton vs Cl… https://t.co/QvvQQUUOG8', retweet_ct=None, retweet=True),
 Row(text='#Uvalde is just another reason why you should home school your kids. Exposing your kids to the degener
ate version o… https://t.co/KiEpSzUJtf', retweet_ct=None, retweet=True),
 Row(text='@TheUSASingers My daughter had a lockdown incident today at her school today hours before what happene
d in Texas😩 M… https://t.co/GdKqjKVjb2', retweet_ct=None, retweet=True),
 Row(text='@iiTorchTv school shooting', retweet_ct=None, retweet=True),
 Row(text='Praying for every demented person who has the guts to pull a trigger and take innocent people’s lives…
pathetic sad… https://t.co/pnbAXP7nqG', retweet_ct=None, retweet=True),
 Row(text='@jersey12jrod @AKurnava @TanjiroSpector @KadenWall11 @ChrisEvans I guarantee there are more people in
Texas conceal… https://t.co/qN4OavtMeQ', retweet_ct=None, retweet=True),
 Row(text='School shooters do NOT get to hijack the struggles of mental health.', retweet_ct=None, retweet=True),
 Row(text='since it\'s about time for graduation, ive seen a few posts saying things along the lines of "i surviv
ed high school… https://t.co/NIUyOkf2tC', retweet_ct=None, retweet=True),
 Row(text="@LynneM021 Heaven forbid one of the Republican's kids was in that school. Wonder what their response w
ould be. Cuz… https://t.co/YMM6s26knU", retweet_ct=None, retweet=True),
 Row(text='Biden Addresses the Nation Tonight About the Texas School Shooting That Killed 14 Children and a\xa0Te
acher https://t.co/Iy7c2rRHuL', retweet_ct=None, retweet=True)]
```

In [37]:

```
new.select('user_loc').groupBy('user_loc').count().orderBy(col('count').desc()).show()
```

```
[Stage 27:================================>                       (11 + 8) / 19]
+---------------+--------+
|       user_loc|   count|
+---------------+--------+
|           null|38274583|
|  United States| 1294357|
| Lagos, Nigeria|  524727|
|        Nigeria|  442125|
|            USA|  421149|
|          India|  385843|
|California, USA|  367685|
|Los Angeles, CA|  352147|
|     Texas, USA|  327748|
|   Florida, USA|  299334|
|London, England|  279549|
|    Chicago, IL|  277230|
|    Houston, TX|  265929|
|         Canada|  260949|
|        she/her|  253710|
| Washington, DC|  253119|
|    Atlanta, GA|  251056|
| United Kingdom|  234331|
|   New York, NY|  211942|
|  New York, USA|  202066|
+---------------+--------+
only showing top 20 rows
```

In [9]:

```
new.write.format("parquet").\
mode('overwrite').\
save('gs://'+'msca-bdp-students-bucket/shared_data/chenfeng/project/eda/')
```

In [67]:

```
place = data.select('place').filter(col("place").isNotNull())
place.withColumn("place", col("place").getItem("name")).limit(100).show()
```

```
+---------------+
|          place|
+---------------+
|         Austin|
|     Enterprise|
|  Warner Robins|
|College Station|
|        Markham|
|       Maryland|
|   South Orange|
|       Hamilton|
|       San Jose|
|    Los Angeles|
|     Rawalpindi|
|        Houston|
|         Austin|
|         Okemos|
|          Texas|
|       Adelaide|
|       Hamilton|
|   Philadelphia|
|    Los Angeles|
|        Florida|
+---------------+
only showing top 20 rows
```

In [64]:

```
place = data.select('place').filter(col("place").isNotNull())
place.withColumn("place", col("place").getItem("country")).limit(100).show()
```

```
+-------------+
|        place|
+-------------+
|United States|
|United States|
|United States|
|United States|
|       Canada|
|United States|
|United States|
|United States|
|United States|
|United States|
|     Pakistan|
|United States|
|United States|
|United States|
|United States|
|    Australia|
|United States|
|United States|
|United States|
|United States|
+-------------+
only showing top 20 rows
```

In [59]:

```
place_count = data.select('place').filter(col("place").isNotNull()).count()
place_count
```

Out[59]:

```
861878
```

In [1]:

```
from pyspark.sql.functions import *
data = spark.read.parquet('gs://'+'msca-bdp-students-bucket/shared_data/chenfeng/project/eda/')
```

In [2]:

```
data.printSchema()
```

```
root
 |-- user_id: long (nullable = true)
 |-- user_name: string (nullable = true)
 |-- user_loc: string (nullable = true)
 |-- user_descrip: string (nullable = true)
 |-- user_followerct: long (nullable = true)
 |-- created_at: string (nullable = true)
 |-- text: string (nullable = true)
 |-- id: long (nullable = true)
 |-- retweeted_from: string (nullable = true)
 |-- retweet_ct: long (nullable = true)
 |-- retweet: boolean (nullable = true)
 |-- full_text: string (nullable = true)
```

In [3]:

```
data = data.withColumn('full_text', coalesce(data.full_text, data.text))
data = data.drop('text')
```

In [10]:

```
data.select('full_text').filter(col('full_text').isNull()).show()
```

```
[Stage 8:=====================================================>(167 + 1) / 168]
```

```
+---------+
|full_text|
+---------+
+---------+
```

Drop Irrelevant Data

In [13]:

```
!pip uninstall nltk -y
!pip install -U nltk
```

```
Found existing installation: nltk 3.7
Uninstalling nltk-3.7:
  Successfully uninstalled nltk-3.7
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the syste
m package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
Collecting nltk
  Using cached nltk-3.7-py3-none-any.whl (1.5 MB)
Requirement already satisfied: regex>=2021.8.3 in /opt/conda/miniconda3/lib/python3.8/site-packages (from nltk) (
2022.10.31)
Requirement already satisfied: joblib in /opt/conda/miniconda3/lib/python3.8/site-packages (from nltk) (1.2.0)
Requirement already satisfied: click in /opt/conda/miniconda3/lib/python3.8/site-packages (from nltk) (7.1.2)
Requirement already satisfied: tqdm in /opt/conda/miniconda3/lib/python3.8/site-packages (from nltk) (4.64.1)
Installing collected packages: nltk
Successfully installed nltk-3.7
WARNING: Running pip as the 'root' user can result in broken permissions and conflicting behaviour with the syste
m package manager. It is recommended to use a virtual environment instead: https://pip.pypa.io/warnings/venv
```

In [4]:

```
import re
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
```

In [5]:

```
text = data.select('full_text')
text = text.rdd.map(lambda x : x['full_text']).filter(lambda x: x is not None)
StopWords = stopwords.words("english")
# remove stop words
tokens = text\
    .map( lambda document: document.strip().lower())\
    .map( lambda document: re.split(" ", document))\
    .map( lambda word: [x for x in word if x.isalnum()])\
    .map( lambda word: [x for x in word if len(x) > 3] )\
    .map( lambda word: [x for x in word if x not in StopWords])
```

In [6]:

```
tokens.take(5)
```

Out[6]:

```
[['trying', 'treat', 'graduate', 'school', 'like', 'undergraduate'],
 ['unironically', 'tweet', 'primary', 'school', 'teacher', 'talking'],
 ['started',
  'today',
  'speaking',
  'journalism',
  'class',
  'talked',
  'importance',
  'voting',
  'civic',
  'engagement'],
 ['year', 'mother', 'college', 'climbing', 'double', 'family', 'tore'],
 ['years', 'advice', 'biology', 'trust', 'refers', 'women']]
```

In [7]:

```python
# web search and eda
desc_token = tokens.flatMap(lambda x: x)\
    .map(lambda x : (x,1))\
    .reduceByKey(lambda x,y: x+y)\
    .map(lambda x : (x[1], x[0]))\
    .sortByKey(ascending = False)
```

In [8]:

```python
desc_token.take(20)
```

```
22/12/01 20:12:56 ERROR org.apache.spark.network.client.TransportClient: Failed to send RPC RPC 79787919353871329
30 to /10.128.0.129:59948: java.nio.channels.ClosedChannelException
java.nio.channels.ClosedChannelException
        at io.netty.channel.AbstractChannel$AbstractUnsafe.newClosedChannelException(AbstractChannel.java:957)
        at io.netty.channel.AbstractChannel$AbstractUnsafe.write(AbstractChannel.java:865)
        at io.netty.channel.DefaultChannelPipeline$HeadContext.write(DefaultChannelPipeline.java:1367)
        at io.netty.channel.AbstractChannelHandlerContext.invokeWrite0(AbstractChannelHandlerContext.java:717)
        at io.netty.channel.AbstractChannelHandlerContext.invokeWriteAndFlush(AbstractChannelHandlerContext.java:
764)
        at io.netty.channel.AbstractChannelHandlerContext$WriteTask.run(AbstractChannelHandlerContext.java:1071)
        at io.netty.util.concurrent.AbstractEventExecutor.safeExecute(AbstractEventExecutor.java:164)
        at io.netty.util.concurrent.SingleThreadEventExecutor.runAllTasks(SingleThreadEventExecutor.java:472)
        at io.netty.channel.nio.NioEventLoop.run(NioEventLoop.java:500)
        at io.netty.util.concurrent.SingleThreadEventExecutor$4.run(SingleThreadEventExecutor.java:989)
        at io.netty.util.internal.ThreadExecutorMap$2.run(ThreadExecutorMap.java:74)
        at io.netty.util.concurrent.FastThreadLocalRunnable.run(FastThreadLocalRunnable.java:30)
        at java.lang.Thread.run(Thread.java:750)
22/12/01 20:12:56 WARN org.apache.spark.storage.BlockManagerMasterEndpoint: Error trying to remove broadcast 5 fr
om block manager BlockManagerId(12, hub-msca-bdp-dphub-students-chenfeng-sw-nhpl.c.msca-bdp-students.internal, 33
387, None)
java.io.IOException: Failed to send RPC RPC 7978791935387132930 to /10.128.0.129:59948: java.nio.channels.ClosedC
hannelException
        at org.apache.spark.network.client.TransportClient$RpcChannelListener.handleFailure(TransportClient.java:
363)
        at org.apache.spark.network.client.TransportClient$StdChannelListener.operationComplete(TransportClient.j
ava:340)
        at io.netty.util.concurrent.DefaultPromise.notifyListener0(DefaultPromise.java:577)
        at io.netty.util.concurrent.DefaultPromise.notifyListenersNow(DefaultPromise.java:551)
        at io.netty.util.concurrent.DefaultPromise.notifyListeners(DefaultPromise.java:490)
        at io.netty.util.concurrent.DefaultPromise.setValue0(DefaultPromise.java:615)
        at io.netty.util.concurrent.DefaultPromise.setFailure0(DefaultPromise.java:608)
        at io.netty.util.concurrent.DefaultPromise.tryFailure(DefaultPromise.java:117)
        at io.netty.channel.AbstractChannel$AbstractUnsafe.safeSetFailure(AbstractChannel.java:993)
        at io.netty.channel.AbstractChannel$AbstractUnsafe.write(AbstractChannel.java:865)
        at io.netty.channel.DefaultChannelPipeline$HeadContext.write(DefaultChannelPipeline.java:1367)
        at io.netty.channel.AbstractChannelHandlerContext.invokeWrite0(AbstractChannelHandlerContext.java:717)
        at io.netty.channel.AbstractChannelHandlerContext.invokeWriteAndFlush(AbstractChannelHandlerContext.java:
764)
        at io.netty.channel.AbstractChannelHandlerContext$WriteTask.run(AbstractChannelHandlerContext.java:1071)
        at io.netty.util.concurrent.AbstractEventExecutor.safeExecute(AbstractEventExecutor.java:164)
        at io.netty.util.concurrent.SingleThreadEventExecutor.runAllTasks(SingleThreadEventExecutor.java:472)
        at io.netty.channel.nio.NioEventLoop.run(NioEventLoop.java:500)
        at io.netty.util.concurrent.SingleThreadEventExecutor$4.run(SingleThreadEventExecutor.java:989)
        at io.netty.util.internal.ThreadExecutorMap$2.run(ThreadExecutorMap.java:74)
        at io.netty.util.concurrent.FastThreadLocalRunnable.run(FastThreadLocalRunnable.java:30)
        at java.lang.Thread.run(Thread.java:750)
Caused by: java.nio.channels.ClosedChannelException
        at io.netty.channel.AbstractChannel$AbstractUnsafe.newClosedChannelException(AbstractChannel.java:957)
        ... 12 more
22/12/01 20:12:56 ERROR org.apache.spark.network.client.TransportResponseHandler: Still have 1 requests outstandi
ng when connection from /10.128.0.132:33156 is closed
22/12/01 20:12:56 ERROR org.apache.spark.network.client.TransportResponseHandler: Still have 1 requests outstandi
ng when connection from /10.128.0.132:33148 is closed
22/12/01 20:12:56 WARN org.apache.spark.storage.BlockManagerMasterEndpoint: Error trying to remove broadcast 4 fr
om block manager BlockManagerId(7, hub-msca-bdp-dphub-students-chenfeng-sw-1t7r.c.msca-bdp-students.internal, 390
67, None)
java.io.IOException: Connection from /10.128.0.132:33156 closed
        at org.apache.spark.network.client.TransportResponseHandler.channelInactive(TransportResponseHandler.java
:146)
        at org.apache.spark.network.server.TransportChannelHandler.channelInactive(TransportChannelHandler.java:1
17)
```

```
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:262)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:248)
        at io.netty.channel.AbstractChannelHandlerContext.fireChannelInactive(AbstractChannelHandlerContext.java:
241)
        at io.netty.channel.ChannelInboundHandlerAdapter.channelInactive(ChannelInboundHandlerAdapter.java:81)
        at io.netty.handler.timeout.IdleStateHandler.channelInactive(IdleStateHandler.java:277)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:262)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:248)
        at io.netty.channel.AbstractChannelHandlerContext.fireChannelInactive(AbstractChannelHandlerContext.java:
241)
        at io.netty.channel.ChannelInboundHandlerAdapter.channelInactive(ChannelInboundHandlerAdapter.java:81)
        at org.apache.spark.network.util.TransportFrameDecoder.channelInactive(TransportFrameDecoder.java:225)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:262)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:248)
        at io.netty.channel.AbstractChannelHandlerContext.fireChannelInactive(AbstractChannelHandlerContext.java:
241)
        at io.netty.channel.DefaultChannelPipeline$HeadContext.channelInactive(DefaultChannelPipeline.java:1405)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:262)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:248)
        at io.netty.channel.DefaultChannelPipeline.fireChannelInactive(DefaultChannelPipeline.java:901)
        at io.netty.channel.AbstractChannel$AbstractUnsafe$8.run(AbstractChannel.java:818)
        at io.netty.util.concurrent.AbstractEventExecutor.safeExecute(AbstractEventExecutor.java:164)
        at io.netty.util.concurrent.SingleThreadEventExecutor.runAllTasks(SingleThreadEventExecutor.java:472)
        at io.netty.channel.nio.NioEventLoop.run(NioEventLoop.java:497)
        at io.netty.util.concurrent.SingleThreadEventExecutor$4.run(SingleThreadEventExecutor.java:989)
        at io.netty.util.internal.ThreadExecutorMap$2.run(ThreadExecutorMap.java:74)
        at io.netty.util.concurrent.FastThreadLocalRunnable.run(FastThreadLocalRunnable.java:30)
        at java.lang.Thread.run(Thread.java:750)
22/12/01 20:12:56 WARN org.apache.spark.storage.BlockManagerMasterEndpoint: Error trying to remove broadcast 4 fr
om block manager BlockManagerId(9, hub-msca-bdp-dphub-students-chenfeng-sw-1t7r.c.msca-bdp-students.internal, 418
55, None)
java.io.IOException: Connection from /10.128.0.132:33148 closed
        at org.apache.spark.network.client.TransportResponseHandler.channelInactive(TransportResponseHandler.java
:146)
        at org.apache.spark.network.server.TransportChannelHandler.channelInactive(TransportChannelHandler.java:1
17)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:262)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:248)
        at io.netty.channel.AbstractChannelHandlerContext.fireChannelInactive(AbstractChannelHandlerContext.java:
241)
        at io.netty.channel.ChannelInboundHandlerAdapter.channelInactive(ChannelInboundHandlerAdapter.java:81)
        at io.netty.handler.timeout.IdleStateHandler.channelInactive(IdleStateHandler.java:277)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:262)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:248)
        at io.netty.channel.AbstractChannelHandlerContext.fireChannelInactive(AbstractChannelHandlerContext.java:
241)
        at io.netty.channel.ChannelInboundHandlerAdapter.channelInactive(ChannelInboundHandlerAdapter.java:81)
        at org.apache.spark.network.util.TransportFrameDecoder.channelInactive(TransportFrameDecoder.java:225)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:262)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:248)
        at io.netty.channel.AbstractChannelHandlerContext.fireChannelInactive(AbstractChannelHandlerContext.java:
241)
        at io.netty.channel.DefaultChannelPipeline$HeadContext.channelInactive(DefaultChannelPipeline.java:1405)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:262)
        at io.netty.channel.AbstractChannelHandlerContext.invokeChannelInactive(AbstractChannelHandlerContext.jav
a:248)
        at io.netty.channel.DefaultChannelPipeline.fireChannelInactive(DefaultChannelPipeline.java:901)
        at io.netty.channel.AbstractChannel$AbstractUnsafe$8.run(AbstractChannel.java:818)
        at io.netty.util.concurrent.AbstractEventExecutor.safeExecute(AbstractEventExecutor.java:164)
        at io.netty.util.concurrent.SingleThreadEventExecutor.runAllTasks(SingleThreadEventExecutor.java:472)
        at io.netty.channel.nio.NioEventLoop.run(NioEventLoop.java:497)
        at io.netty.util.concurrent.SingleThreadEventExecutor$4.run(SingleThreadEventExecutor.java:989)
        at io.netty.util.internal.ThreadExecutorMap$2.run(ThreadExecutorMap.java:74)
        at io.netty.util.concurrent.FastThreadLocalRunnable.run(FastThreadLocalRunnable.java:30)
        at java.lang.Thread.run(Thread.java:750)
```

Out[8]:

```
[(39735231, 'school'),
 (11044189, 'college'),
 (8461610, 'high'),
 (8139660, 'university'),
 (6468330, 'schools'),
 (4835983, 'like'),
 (4739795, 'students'),
 (3665067, 'kids'),
 (3507005, 'people'),
 (2915399, 'professor'),
 (2876460, 'back'),
 (2824178, 'children'),
 (2717518, 'would'),
 (2633895, 'first'),
 (2602132, 'time'),
 (2600443, 'student'),
 (2581039, 'public'),
 (2487809, 'year'),
 (2425646, 'know'),
 (2260432, 'want')]
```

In [24]:

```
keywords = ['high', 'college', 'university', 'secondary', 'primary', 'education', 'k-12', 'professor', 'students'
            'kids', 'school', 'schools', 'graduate', 'undergraduate', 'kids', 'children', 'public', 'private']
```

In [23]:

```
tokens.zip(data.select('id').rdd.flatMap(lambda x:x)).take(5)
```

Out[23]:

```
[(['trying', 'treat', 'graduate', 'school', 'like', 'undergraduate'],
  1571982079197581313),
 (['unironically', 'tweet', 'primary', 'school', 'teacher', 'talking'],
  1571982082343112705),
 (['started',
   'today',
   'speaking',
   'journalism',
   'class',
   'talked',
   'importance',
   'voting',
   'civic',
   'engagement'],
  1571982082305593345),
 (['year', 'mother', 'college', 'climbing', 'double', 'family', 'tore'],
  1571982083027005441),
 (['years', 'advice', 'biology', 'trust', 'refers', 'women'],
  1571982083140235266)]
```

In [6]:

```
from pyspark.sql.types import  StructType, StructField, ArrayType, IntegerType
from pyspark.sql import Row
t = tokens.zip(data.select('id').rdd.flatMap(lambda x:x))
df_tokens = t.toDF(['tokens', 'id'])
df_tokens.printSchema()
```

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

root
 |-- tokens: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- id: long (nullable = true)
```

In [64]:

```
df_tokens.show(5)
```

```
+--------------------+-------------------+
|              tokens|                 id|
+--------------------+-------------------+
|[trying, treat, g...|1571982079197581313|
|[unironically, tw...|1571982082343112705|
|[started, today, ...|1571982082305593345|
|[year, mother, co...|1571982083027005441|
|[years, advice, b...|1571982083140235266|
+--------------------+-------------------+
only showing top 5 rows
```

In [68]:

```
d = df_tokens.limit(5)
d.show()
```

```
[Stage 48:=====================================================>(291 + 2) / 293]
+--------------------+-------------------+
|              tokens|                 id|
+--------------------+-------------------+
|[solve, think, co...|1580989648968372224|
|[stop, living, li...|1580989649635246083|
|[favorite, part, ...|1580989649820225537|
|[listening, colle...|1580989649849180160|
|[feel, like, talk...|1580989650331521024|
+--------------------+-------------------+
```

In [69]:

```
d.collect()
```

Out[69]:

```
[Row(tokens=['trying', 'treat', 'graduate', 'school', 'like', 'undergraduate'], id=1571982079197581313),
 Row(tokens=['unironically', 'tweet', 'primary', 'school', 'teacher', 'talking'], id=1571982082343112705),
 Row(tokens=['started', 'today', 'speaking', 'journalism', 'class', 'talked', 'importance', 'voting', 'civic', 'e
ngagement'], id=1571982082305593345),
 Row(tokens=['year', 'mother', 'college', 'climbing', 'double', 'family', 'tore'], id=1571982083027005441),
 Row(tokens=['years', 'advice', 'biology', 'trust', 'refers', 'women'], id=1571982083140235266)]
```

In [75]:

```
d.cache()
```

Out[75]:

```
DataFrame[tokens: array<string>, id: bigint]
```

In [77]:

```
d.show()
```

```
+--------------------+-------------------+
|              tokens|                 id|
+--------------------+-------------------+
|[trying, treat, g...|1571982079197581313|
|[unironically, tw...|1571982082343112705|
|[started, today, ...|1571982082305593345|
|[year, mother, co...|1571982083027005441|
|[years, advice, b...|1571982083140235266|
+--------------------+-------------------+
```

In [82]:

```
d.write.saveAsTable('dt')
```

```
22/12/01 23:35:30 WARN org.apache.hadoop.hive.ql.session.SessionState: METASTORE_FILTER_HOOK will be ignored, sin
ce hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.
```

In [84]:

```
spark.sql("select tokens from dt").show()
```

```
+--------------------+
|              tokens|
+--------------------+
|[trying, treat, g...|
|[unironically, tw...|
|[started, today, ...|
|[year, mother, co...|
|[years, advice, b...|
+--------------------+
```

In [88]:

```
query = "select * from dt where array_contains (tokens, 'year') \
        or array_contains (tokens, 'years')"
spark.sql(query).show()
```

```
+--------------------+-------------------+
|              tokens|                 id|
+--------------------+-------------------+
|[year, mother, co...|1571982083027005441|
|[years, advice, b...|1571982083140235266|
+--------------------+-------------------+
```

In [ ]:

```
keywords = ['high', 'college', 'university', 'secondary', 'primary', 'education', 'k-12', 'professor', 'students'
            'kids', 'school', 'schools', 'graduate', 'undergraduate', 'kids', 'children', 'public', 'private','te
```

In [7]:

```
df_tokens.write.saveAsTable('dtt')
```

```
ivysettings.xml file not found in HIVE_HOME or HIVE_CONF_DIR,/etc/hive/conf.dist/ivysettings.xml will be used
22/12/04 04:11:03 WARN org.apache.hadoop.hive.ql.session.SessionState: METASTORE_FILTER_HOOK will be ignored, sin
ce hive.security.authorization.manager is set to instance of HiveAuthorizerFactory.
```

In [8]:

```
query = """
        select * from dtt where array_contains (tokens, 'high') \
        or array_contains (tokens, 'college') \
        or array_contains (tokens, 'university') \
        or array_contains (tokens, 'secondary') \
        or array_contains (tokens, 'primary') \
        or array_contains (tokens, 'education') \
        or array_contains (tokens, 'k-12') \
        or array_contains (tokens, 'undergraduate') \
        or array_contains (tokens, 'graduate')
        """
```

In [9]:

```
key_df = spark.sql(query)
```

In [152…

```
key_df.show()
```

```
+--------------------+-------------------+
|              tokens|                 id|
+--------------------+-------------------+
|[trying, treat, g...|1571982079197581313|
|[unironically, tw...|1571982082343112705|
|[year, mother, co...|1571982083027005441|
|[elder, knows, st...|1571982084301852675|
|[herschel, high, ...|1571982090572541952|
|[□□□□□, GUNN, ...|1571982090765307906|
|[part, college, s...|1571982091038126081|
|           [college]|1571982091016933377|
|[rush, limbaugh, ...|1571982091323064321|
|[2020, trump, pla...|1571982091784536064|
|     [hate, college]|1571982092464160768|
|[sleepers, upcomi...|1571982092757794816|
|[appreciate, sinc...|1571982093856673792|
|[university, brid...|1571982098910842880|
|[coastal, elite, ...|1571982100089212933|
|[work, federal, c...|1571982102769319936|
|[football, high, ...|1571982107219574784|
|[kirtland, geneva...|1571982108083777537|
|[college, taught,...|1571982112504385540|
|[posted, photo, s...|1571982113586520068|
+--------------------+-------------------+
only showing top 20 rows
```

In [10]:

```
df = key_df.join(data, key_df.id == data.id, 'inner').drop(key_df['id'])
```

In [144…

```
df.show(10)
```

```
[Stage 87:>                                                         (0 + 1) / 1]
+--------------------+-------------------+---------------+--------------------+------------+---------------+-----
---------------+-------------------+--------------+----------+-------+--------------------+
|              tokens|            user_id|      user_name|            user_loc|user_favorct|user_followerct|
created_at|                 id|retweeted_from|retweet_ct|retweet|           full_text|
+--------------------+-------------------+---------------+--------------------+------------+---------------+-----
---------------+-------------------+--------------+----------+-------+--------------------+
|[flame, soon, lea...|         2442064466|      p_rosepro|         Chicago, IL|       20049|           1060|Tue A
pr 05 04:22:...|1511197458021421059|          null|      null|   null|I flame up as soo...|
|[want, something,...|1315736023566090242|         g00ger|hammond IN / chicago|        7405|            144|Tue A
pr 05 04:23:...|1511197924092522501|          null|      null|   null|@coharesexcface I...|
|[holy, issued, pr...|           17677908|JanakinathSahay|Indraprastha & Pr...|       19365|            412|Tue A
pr 05 04:24:...|1511198100366364672|  libsoftiktok|       843|  false|RT @libsoftiktok:...|
|[tottenham, 30yrs...|          904896402|     spurfect82|                null|        5488|             71|Tue A
pr 05 04:25:...|1511198327429382147|          null|      null|   null|@talkSPORT  I’m a...|
|[glad, taking, br...| 980626337139392512| victoriaboulom| asian, laotian | 16|        9239|             34|Tue A
pr 05 04:25:...|1511198383507079180|          null|      null|   null|i’m so glad i’m t...|
|[mandatory, obtai...|1208279755360854016|Krishan13399026|      Gurgaon, India|        6403|             87|Tue A
pr 05 04:28:...|1511198964497977349|          null|      null|   null|It's mandatory to...|
|[thank, college, ...| 764170471197798400|   sellthedolan|       New York, USA|        6808|            566|Tue A
pr 05 04:31:...|1511199739181776901|  bubbagumpino|        25|  false|RT @bubbagumpino:...|
|[proud, moment, l...|          786237384|Rahultahiliani9|   Chandigarh, India|        4851|           4575|Tue A
pr 05 04:31:...|1511199808144678915| raghav_chadha|         8|  false|RT @raghav_chadha...|
|[hope, college, c...|1023813199387402240|      yoooshay1|                null|        4980|            328|Tue A
pr 05 04:32:...|1511200049761501185|   archiedgzmn|      8250|  false|RT @archiedgzmn: ...|
|[ever, tell, favo...|           13447022|        sdaoudi|         Olympia, WA|        6347|            522|Tue A
pr 05 04:33:...|1511200218292756484|          null|      null|   null|@donventura Did I...|
+--------------------+-------------------+---------------+--------------------+------------+---------------+-----
---------------+-------------------+--------------+----------+-------+--------------------+
only showing top 10 rows
```

In [11]:

```
df.write.format("parquet").\
mode('overwrite').\
save('gs://'+'msca-bdp-students-bucket/shared_data/chenfeng/project/key_eda/')
```

In [12]:

```
df.count()
```

Out[12]:

```
29128862
```

In [146…

```
data.count()
```

Out[146…

```
99992797
```

In [156…

```
df.printSchema()
```

```
root
 |-- tokens: array (nullable = true)
 |    |-- element: string (containsNull = true)
 |-- user_id: long (nullable = true)
 |-- user_name: string (nullable = true)
 |-- user_loc: string (nullable = true)
 |-- user_favorct: long (nullable = true)
 |-- user_followerct: long (nullable = true)
 |-- created_at: string (nullable = true)
 |-- id: long (nullable = true)
 |-- retweeted_from: string (nullable = true)
 |-- retweet_ct: long (nullable = true)
 |-- retweet: boolean (nullable = true)
 |-- full_text: string (nullable = true)
```

In [ ]:

```

```

In [ ]:

```

```