# Influencer

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
from pyspark.sql.functions import *
text = spark.read.parquet('gs://'+'msca-bdp-students-bucket/shared_data/chenfeng/project/key_eda/')
```

In [10]:
```
text.show(5)
```

```
[Stage 24:>                                                         (0 + 1) / 1]
+--------------------+------------------+-------------+----------------+--------------------+--------------+---
-----------------+------------------+-------------+----------+-------+--------------------+
|              tokens|           user_id|    user_name|        user_loc|        user_descrip|user_followerct|
created_at|                id|retweeted_from|retweet_ct|retweet|           full_text|
+--------------------+------------------+-------------+----------------+--------------------+--------------+---
-----------------+------------------+-------------+----------+-------+--------------------+
|[sometimes, dysph...|1511006361517117446| mroberholt__|    Bay Area, CA|I am here b/c the...|             0|Tue
Apr 05 04:23:...|1511197879397806084|         null|      null|   null|I sometimes get d...|
|[shoutout, fellow...| 825962906659454976|setherfan3244|  Mile High City|Dog and cat mom. ...|           117|Tue
Apr 05 04:24:...|1511198184298586115|   SheaSerrano|       194|  false|RT @SheaSerrano: ...|
|[semester, taught...|1024768034655887364| nia_simone27|   United States|ATL ✈  DMV || How...|          1153|Tue
Apr 05 04:26:...|1511198670347292672|    akume_oben|        27|  false|RT @akume_oben: T...|
|[1988, local, uni...|         109060885|Sidearm2Kelce|         Georgia| Lifelong sports fan|           279|Tue
Apr 05 04:30:...|1511199542934491140|   princetongb|         1|  false|RT @princetongb: ...|
|[delegation, liro...|        3314963708|    Bhaweshkj|New Delhi, India|STPI, Govt. Of In...|           281|Tue
Apr 05 04:32:...|1511200011505532928|     stpiindia|       188|  false|RT @stpiindia: De...|
+--------------------+------------------+-------------+----------------+--------------------+--------------+---
-----------------+------------------+-------------+----------+-------+--------------------+
only showing top 5 rows
```

In [5]:
```
original = text.filter(col('retweet').isNull())
retweet = text.filter(col('retweet').isNotNull())
```

## Volume of original content

In [17]:
```
ori_ct = original.groupby('user_id').count().sort(col('count').desc())
ori_ct.show()
```

```
[Stage 29:=========================================>               (5 + 2) / 7]
+------------------+-----+
|           user_id|count|
+------------------+-----+
|1463182041147576321|19712|
|1422259384525090818|19679|
|1483749445900861446|12649|
|1128225338775953408|11725|
|         219401992| 9236|
| 879496394691805184| 8742|
|1426164518581899266| 8638|
|1322084295507259392| 8498|
|1508968207259869185| 8488|
|1323262598591074307| 8319|
|1321667597469769728| 8161|
|1434562949381767177| 7691|
|1338108136561856517| 7329|
|1496554077349683207| 7257|
|1300833220661977088| 6683|
|1300839891769589760| 6441|
|         2538008610| 6436|
|1329639203144011776| 6433|
|1323137083335045121| 6134|
|         1079741329| 6075|
```

```
+-------------------+-----+
only showing top 20 rows
```

```
22/12/02 05:21:36 WARN org.apache.spark.deploy.yarn.YarnAllocator: Container from a bad node: container_166969513
9548_0016_01_000018 on host: hub-msca-bdp-dphub-students-chenfeng-sw-zqc9.c.msca-bdp-students.internal. Exit stat
us: 143. Diagnostics: [2022-12-02 05:21:36.256]Container killed on request. Exit code is 143
[2022-12-02 05:21:36.257]Container exited with a non-zero exit code 143.
[2022-12-02 05:21:36.257]Killed by external signal
.
22/12/02 05:21:36 WARN org.apache.spark.deploy.yarn.YarnAllocator: Container from a bad node: container_166969513
9548_0016_01_000016 on host: hub-msca-bdp-dphub-students-chenfeng-sw-zqc9.c.msca-bdp-students.internal. Exit stat
us: 143. Diagnostics: [2022-12-02 05:21:36.256]Container killed on request. Exit code is 143
[2022-12-02 05:21:36.256]Container exited with a non-zero exit code 143.
[2022-12-02 05:21:36.257]Killed by external signal
.
22/12/02 05:21:36 ERROR org.apache.spark.scheduler.cluster.YarnScheduler: Lost executor 17 on hub-msca-bdp-dphub-
students-chenfeng-sw-zqc9.c.msca-bdp-students.internal: Container from a bad node: container_1669695139548_0016_0
1_000018 on host: hub-msca-bdp-dphub-students-chenfeng-sw-zqc9.c.msca-bdp-students.internal. Exit status: 143. Di
agnostics: [2022-12-02 05:21:36.256]Container killed on request. Exit code is 143
[2022-12-02 05:21:36.257]Container exited with a non-zero exit code 143.
[2022-12-02 05:21:36.257]Killed by external signal
.
22/12/02 05:21:36 WARN org.apache.spark.scheduler.cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Requesting
driver to remove executor 17 for reason Container from a bad node: container_1669695139548_0016_01_000018 on host
: hub-msca-bdp-dphub-students-chenfeng-sw-zqc9.c.msca-bdp-students.internal. Exit status: 143. Diagnostics: [2022
-12-02 05:21:36.256]Container killed on request. Exit code is 143
[2022-12-02 05:21:36.257]Container exited with a non-zero exit code 143.
[2022-12-02 05:21:36.257]Killed by external signal
.
22/12/02 05:21:36 WARN org.apache.spark.scheduler.cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Requesting
driver to remove executor 15 for reason Container from a bad node: container_1669695139548_0016_01_000016 on host
: hub-msca-bdp-dphub-students-chenfeng-sw-zqc9.c.msca-bdp-students.internal. Exit status: 143. Diagnostics: [2022
-12-02 05:21:36.256]Container killed on request. Exit code is 143
[2022-12-02 05:21:36.256]Container exited with a non-zero exit code 143.
[2022-12-02 05:21:36.257]Killed by external signal
.
22/12/02 05:21:36 ERROR org.apache.spark.scheduler.cluster.YarnScheduler: Lost executor 15 on hub-msca-bdp-dphub-
students-chenfeng-sw-zqc9.c.msca-bdp-students.internal: Container from a bad node: container_1669695139548_0016_0
1_000016 on host: hub-msca-bdp-dphub-students-chenfeng-sw-zqc9.c.msca-bdp-students.internal. Exit status: 143. Di
agnostics: [2022-12-02 05:21:36.256]Container killed on request. Exit code is 143
[2022-12-02 05:21:36.256]Container exited with a non-zero exit code 143.
[2022-12-02 05:21:36.257]Killed by external signal
.
```

In [21]:

```
original.filter(col('user_id') == 1463182041147576321).\
select('user_id', 'user_name', 'user_loc', 'user_followerct','full_text').show()
```

```
+-------------------+---------+-------------+---------------+--------------------+
|            user_id|user_name|     user_loc|user_followerct|           full_text|
+-------------------+---------+-------------+---------------+--------------------+
|1463182041147576321|sport9920|united states|             27|[LIVE] Notre Dame...|
|1463182041147576321|sport9920|united states|             31|[LIVE] Calvary Ch...|
|1463182041147576321|sport9920|united states|             35|[LIVE] Perham vs....|
|1463182041147576321|sport9920|united states|             35|: Tonight's @ 5...|
|1463182041147576321|sport9920|united states|             34|LIVE ~ Southsid...|
|1463182041147576321|sport9920|united states|             34|LIVE ~ Penobsco...|
|1463182041147576321|sport9920|united states|             34|Silver Lake Regio...|
|1463182041147576321|sport9920|united states|             36|Fallbrook High Sc...|
|1463182041147576321|sport9920|united states|             36|Cheshire High Sch...|
|1463182041147576321|sport9920|united states|             36|Darien High Schoo...|
|1463182041147576321|sport9920|united states|             36|Severna Park High...|
|1463182041147576321|sport9920|united states|             36| Live▶ Telstar ...|
|1463182041147576321|sport9920|united states|             40|Live▶ Caro High...|
|1463182041147576321|sport9920|united states|             41|: Tonight's @ 7...|
|1463182041147576321|sport9920|united states|             41|[LIVE] Cochranton...|
|1463182041147576321|sport9920|united states|             41|Live▶ Central S...|
|1463182041147576321|sport9920|united states|             44|: Tonight's @ 8...|
|1463182041147576321|sport9920|united states|             42|Glen Allen High S...|
|1463182041147576321|sport9920|united states|             44|: Tonight's @ 7...|
|1463182041147576321|sport9920|united states|             44|Live▶ Greendale...|
+-------------------+---------+-------------+---------------+--------------------+
only showing top 20 rows
```

In [20]:
```
ori_followerct = original.groupby('user_id').agg(max('user_followerct').alias('followerct'))
ori_followerct.show()
```

```
[Stage 36:=====================================================>  (48 + 2) / 50]
```

```
+------------------+----------+
|           user_id|followerct|
+------------------+----------+
|1251492751293775872|        28|
|1310622940971651074|         5|
|          52769844|       415|
|         445881364|      1319|
|1167204364622151685|         3|
|         270182627|       700|
|          17028478|      9988|
|        2415993637|        56|
|        1574062046|       474|
|          22146921|      1417|
|         243813195|      1504|
|          49097621|      1337|
|         951875166|       395|
|          14194973|      7379|
|        2262451992|      1114|
|1464223636831805476|       196|
|        1676058542|       136|
|         279598130|       439|
|         127038335|      1478|
|         125719746|       782|
+------------------+----------+
only showing top 20 rows
```

In [19]:
```
ori_name = original.groupby('user_id').agg(max('user_name').alias('user_name'))
ori_name.show()
```

```
[Stage 33:=====================================================> (49 + 1) / 50]
+-------+------------+
|user_id|   user_name|
+-------+------------+
|     22|      rabble|
|    418|        dens|
|   1427|   tchambers|
|   1497|      harper|
|   2489|    psulli21|
|   2590|  dangrsmind|
|   2900| jeffisageek|
|   3252|   odannyboy|
|   4509|andyabramson|
|   4603|       jem27|
|   4677|      barryf|
|   4999|     fraying|
|   5279|   moonhouse|
|   5605| msgenevieve|
|   5618|  sharpshoot|
|   5628|         flc|
|   6136|aotearoa_ben|
|   6378|       aDeSe|
|   8264|      alicia|
|   8412| alexhillman|
+-------+------------+
only showing top 20 rows
```

In [27]:
```
user = ori_ct.join(ori_name, ori_ct.user_id == ori_name.user_id, 'inner').drop(ori_name['user_id'])
```

In [28]:
```
user.show()
```

[Stage 60:> (0 + 1) / 1]

```
+-----+-------+-------------+
|count|user_id|    user_name|
+-----+-------+-------------+
|    1|     22|       rabble|
|    1|    418|         dens|
|    1|   1427|    tchambers|
|    4|   1497|       harper|
|    1|   2294|         tito|
|    3|   2360|anshulkundaje|
|    1|   2489|     psulli21|
|    2|   2590|   dangrsmind|
|    1|   2900|  jeffisageek|
|    1|   3252|    odannyboy|
|    3|   4509| andyabramson|
|    1|   4603|        jem27|
|    1|   4677|       barryf|
|    2|   4999|      fraying|
|    4|   5195|       0xFlip|
|    1|   5279|    moonhouse|
|    3|   5505|      shaycam|
|    1|   5605|  msgenevieve|
|    2|   5618|   sharpshoot|
|    2|   5628|          flc|
+-----+-------+-------------+
only showing top 20 rows
```

In [30]:

```
user_df = user.join(ori_followerct, user.user_id == ori_followerct.user_id, 'inner').drop(ori_followerct['user_id
user_df.show()
```

[Stage 67:> (0 + 1) / 1]

```
+-----+-------+------------+----------+
|count|user_id|   user_name|followerct|
+-----+-------+------------+----------+
|    1|     22|      rabble|     17736|
|    1|    418|        dens|     83191|
|    1|   1427|   tchambers|      2112|
|    4|   1497|      harper|     38085|
|    1|   2489|    psulli21|        26|
|    2|   2590|  dangrsmind|      1453|
|    1|   2900| jeffisageek|      8083|
|    1|   3252|   odannyboy|     31742|
|    3|   4509|andyabramson|      6379|
|    1|   4603|       jem27|       123|
|    1|   4677|      barryf|       711|
|    2|   4999|     fraying|     10563|
|    1|   5279|   moonhouse|      1666|
|    1|   5605| msgenevieve|     15544|
|    2|   5618|  sharpshoot|      4270|
|    2|   5628|         flc|       428|
|    2|   6136|aotearoa_ben|      9667|
|    1|   6378|       aDeSe|      3593|
|    3|   8264|      alicia|      4879|
|    2|   8412| alexhillman|     13929|
+-----+-------+------------+----------+
only showing top 20 rows
```

In [31]:

```
user_df = user_df.sort(col('count').desc())
user_df.show()
```

[Stage 74:==========================================> (7 + 2) / 9]

```
+-----+------------------+--------------+----------+
|count|           user_id|     user_name|followerct|
+-----+------------------+--------------+----------+
|19712|1463182041147576321|     sport9920|        94|
|19679|1422259384525090818|   ana92479235|       139|
```

```
|12649|1483749445900861446|       AgiwaraS|        43|
|11725|1128225338775953408| AndrianyRahmah|       218|
| 9236|          219401992|  DennisStemmle|      3804|
| 8742| 879496394691805184|   EssayPaperUK|       804|
| 8638|1426164518581899266|     iskolworks|       695|
| 8498|1322084295507259392|       sorth_me|       118|
| 8488|1508968207259869185|        hilmsit|       353|
| 8319|1323262598591074307|   aljunasghost|        66|
| 8161|1321667597469769728|     AswenDiana|        38|
| 7691|1434562949381767177|       Porxlek1|         0|
| 7329|1338108136561856517|  adeliasari033|        21|
| 7257|1496554077349683207|   TheCurrentSa|        18|
| 6683|1300833220661977088|cintia238276122|        69|
| 6441|1300839891769589760|SportsC62770366|        73|
| 6436|         2538008610|         shrs79|        64|
| 6433|1329639203144011776| robson89148161|        30|
| 6134|1323137083335045121|astro090_plo182|        15|
| 6075|         1079741329|     mbahucup84|        31|
+-----+-------------------+---------------+----------+
only showing top 20 rows
```

## Volume of retweet

In [4]:
```
ret_ct = retweet.groupby('id').agg(max('retweet_ct').alias('retweet_ct'))
ret_ct.show()
```

```
[Stage 4:>                                                          (0 + 1) / 1]
+-------------------+----------+
|                 id|retweet_ct|
+-------------------+----------+
|1511375752956358668|       171|
|1512089526004432915|        11|
|1512145944078274578|        81|
|1512146255543214105|        28|
|1512170567738163228|         1|
|1512172627875430425|         1|
|1512204990609207331|     11013|
|1512221798737358879|        39|
|1512223249232539680|         3|
|1512227436959871001|       155|
|1512235002238578716|      5942|
|1512238594911612982|        48|
|1512267294453420086|         1|
|1512267702928298034|      6672|
|1513500420584673296|        42|
|1513674830477938729|       201|
|1514353542915674146|         8|
|1514603291375677489|        13|
|1514608243439378452|      1028|
|1514663529705447446|        64|
+-------------------+----------+
only showing top 20 rows
```

In [17]:
```
user_df.printSchema()
```

```
root
 |-- count: long (nullable = false)
 |-- user_id: long (nullable = true)
 |-- user_name: string (nullable = true)
 |-- followerct: long (nullable = true)
 |-- user_descrip: string (nullable = true)
```

In [12]:
```
ret_user = ret_ct.join(retweet, ret_ct.id==retweet.id,'inner').select(ret_ct.id,ret_ct.retweet_ct,'retweeted_from
ret_user_sum = ret_user.groupby('retweeted_from').agg(sum('retweet_ct').alias('total_rct'))
ret_user_sum = ret_user_sum.sort(col('total_rct').desc())
ret_user_sum.show()
```

```
[Stage 41:>                                                         (0 + 2) / 3]
```

```
+---------------+----------+
| retweeted_from| total_rct|
+---------------+----------+
|     PEScorpiio|2189140651|
| ChrChristensen|1242970553|
|         brndxq| 893227088|
| polevaultpower| 748415520|
|        mattxiv| 648079976|
|MichaelWarbur17| 552349546|
|   Noorthevirgo| 422953683|
|        Jch_210| 390358559|
|   davidhogg111| 379496866|
|     stuckiny2k| 368907377|
|   etanthomas36| 345457606|
|OccupyDemocrats| 341651108|
|     jasminn_02| 314812670|
|        schwarz| 299899941|
|   zandermoricz| 291449083|
| CamiOfahengaue| 289037071|
| _fashionkilaa_| 275045102|
|ellewasamistake| 274488796|
|  peterframpton| 268572754|
|     rhiankatie| 261266852|
+---------------+----------+
only showing top 20 rows
```

In [55]:
```
user_rtc.printSchema()
```

```
root
 |-- count: long (nullable = false)
 |-- user_id: long (nullable = true)
 |-- user_name: string (nullable = true)
 |-- followerct: long (nullable = true)
 |-- user_descrip: string (nullable = true)
 |-- total_rct: long (nullable = true)
```

In [14]:
```
user_rtc = user_df.join(ret_user_sum, user_df.user_name == ret_user_sum.retweeted_from, 'inner').\
drop(ret_user_sum['retweeted_from'])
user_rtc.show()
```

```
[Stage 81:>                                                         (0 + 1) / 1]
+-----+-------------------+---------------+----------+---------+
|count|            user_id|      user_name|followerct|total_rct|
+-----+-------------------+---------------+----------+---------+
|    4|          276525537|         007mss|      3891|       11|
|    3|1263020081275166720|00HarshitMishra|       662|        1|
|    1|          547224666|      01975Bill|        24|        2|
|    1|         2956750401|       01Khan_M|      1221|        2|
|    1|1447623365364224007|01Slut_Princess|      7692|      129|
|    8|          104865102|       01Waller|      1032|        5|
|   17|1518991522041053186|    02091942Jay|      6939|       49|
|    1|1561831078138204161|       024files|        26|        1|
|    1|1308241569624006656|   02brooklyn_a|        53|        9|
|    4|1213348873105657856|02hwangwatanabe|      1636|       95|
|    3|1254578337386725378|      0323x0710|       469|     6385|
|    1|1466284086750707721|       0403FEED|      6814|    15209|
|    2|1397067211039068162|       0423kart|      1582|      124|
|    2|1315371222310572032|05ImpactSpencer|        79|        2|
|    1|1024693018828435456|         05Ohio|       516|        3|
|   11|1277596222820409345|        05Uzair|       956|       58|
|    1|1441161312260669444|   061313030419|       275|        1|
|    2|1483459438824591368|      06Midwest|        36|        1|
|    2|1424257095252078597|       0702cart|       891|       13|
|    5|1334805645090934785|        07Eldho|      1695|       54|
```

```
+-----+------------------+--------------+----------+---------+
only showing top 20 rows
```

In [ ]:

```
#### reproduce section
# original
ori_ct = original.groupby('user_id').count().sort(col('count').desc())
ori_followerct = original.groupby('user_id').agg(max('user_followerct').alias('followerct'))
ori_name = original.groupby('user_id').agg(max('user_name').alias('user_name'))
ori_des = original.groupby('user_id').agg(max('user_descrip').alias('user_descrip'))
user = ori_ct.join(ori_name, ori_ct.user_id == ori_name.user_id, 'inner').drop(ori_name['user_id'])
user_df = user.join(ori_followerct, user.user_id == ori_followerct.user_id, 'inner').drop(ori_followerct['user_id
user_df = user_df.join(ori_des, user.user_id == ori_des.user_id, 'inner').drop(ori_des['user_id'])

# retweet
ret_ct = retweet.groupby('full_text').agg(max('retweet_ct').alias('retweet_ct'))
ret_user = ret_ct.join(retweet, ret_ct.full_text==retweet.full_text,'left').\
select(ret_ct.full_text,ret_ct.retweet_ct,'retweeted_from')
ret_user = ret_user.dropDuplicates()
ret_user_sum = ret_user.groupby('retweeted_from').agg(sum('retweet_ct').alias('total_rct'))
user_rtc = user_df.join(ret_user_sum, user_df.user_name == ret_user_sum.retweeted_from, 'inner').\
drop(ret_user_sum['retweeted_from'])
```

In [ ]:

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

In [68]:

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

In [59]:

```
inf.show()
```

```
+-----+------------------+-------------+----------+----------------------+---------+
|count|           user_id|    user_name|followerct|          user_descrip|total_rct|
+-----+------------------+-------------+----------+----------------------+---------+
|    1|         129963753|   000ricardo|       193|  PhD Candidate in ...|        1|
|    3|1057645964763283457| 001Legendary|      2359|  Tall || Unilorite...|        4|
|    2|        3023363029|     007DABBS|       734|       †  0930.7EVEN|        1|
|    1|        1228879926|   00honeybee|       931|  Anti-Communist. I...|        1|
|    2|         134993455| 00princejeno|      6204|  210123 #JENO: bab...|        4|
|    1|1236245494600368129|       01Semu|       831|  Working  hard doe...|        4|
|    2|1522102125123485696|     01dGuard|       885|  travelling mostly...|     1065|
|    1|1367688645457698819|        01x1x|       378|  Humare Yaha libra...|        2|
|   38|         230736536|       020644|       149|                  null|        1|
|    2|1585905427111452672|    02_Misbah|         2|                  null|      955|
|    3|1058878571454832640|     0325skzz|       381|  @yunhoeshi on twi...|        1|
|    1|1432521098474790914|   03_rayosfc|        29|  2003/04 Boys comp...|        1|
|    3| 876384543954972672|        03fma|       183|20↑. genshin. fma▶ ...|        1|
|    8|1352312678698647552|     0402Mads|       869|  Madison H - C/o 2...|        2|
|    2|1444277784310812678|04BottleVodka|       838|  Indian □□
Nothi...|       24|
|    1|1308423796039221249|    04Sisco40|       313|  •Francisco•any pr...|        8|
|    1|         390685741|      054Igwe|       831|  ✝  Christian | □...|        1|
|    1|1131244984911503360| 05NCFCELITEG|       138|  2022-2023 Twitter...|        1|
|    3|1409623145380732935|   0613frames|     23555|  For cinema, media...|    12520|
|    1|         2549299226|    066025SSS|       610|  An Engineer in L&...|        1|
+-----+------------------+-------------+----------+----------------------+---------+
only showing top 20 rows
```

In [2]:

```
from pyspark.sql import functions as F
from pyspark.sql import types as t
from pyspark.sql.types import ArrayType, IntegerType, BooleanType

inf = inf.filter(col('user_descrip').isNotNull())
eng_ord=F.udf(lambda x: [ord(a) for a in x],t.ArrayType(IntegerType()))
def eng_filter(x):
    for index in range(len(x)):
        if x[index] > 128:
            return False
    return True
```

```
filter_udf = F.udf(eng_filter, BooleanType())
inf = inf.filter(filter_udf(eng_ord(col('user_descrip'))) == True)
```

In [61]:
```
inf.show()
```

```
[Stage 94:>                                                         (0 + 1) / 1]
+-----+------------------+-------------+----------+--------------------+---------+
|count|           user_id|    user_name|followerct|        user_descrip|total_rct|
+-----+------------------+-------------+----------+--------------------+---------+
|    1|         129963753|   000ricardo|       193|PhD Candidate in ...|        1|
|    1|1236245494600368129|       01Semu|       831|Working  hard doe...|        4|
|    2|1522102125123485696|     01dGuard|       885|travelling mostly...|     1065|
|    3|1058878571454832640|     0325skzz|       381|@yunhoeshi on twi...|        1|
|    1|1432521098474790914|   03_rayosfc|        29|2003/04 Boys comp...|        1|
|    1|1131244984911503360| 05NCFCELITEG|       138|2022-2023 Twitter...|        1|
|    3|1409623145380732935|   0613frames|     23555|For cinema, media...|    12520|
|    1|         2549299226|    066025SSS|       610|An Engineer in L&...|        1|
|    1|1546187325624967169|       06gfdl|       100|Team of girls pla...|        1|
|    6|         2359329387|    07003ARTS|       744|Jenn Khoury| Dist...|        5|
|    1|           25119095| 07003bhsband|       359|Bloomfield High S...|        3|
|    1|1293296242143961091|      07gDksc|        99|The Official DKSC...|       20|
|    1|         1542528403| 0BurkeBlack0|     30452|Pirate Captain ou...|        2|
|    4|1349129808710688770|    0IuwaFEMl|      2920|father and husban...|    68345|
|    2|1343102471770697729| 0SAMUTIDDIES|     11753|ushisakuatsu enjo...|  1176664|
|    1|         2185503830|    0_H_I_h0e|      1176|Definitely a huma...|        6|
|   17|1468314153467006977|0detteroulette|      1107|                  no|        9|
|    8|          392472061|     0hbetave|      1022|"I just want the ...|      123|
|    6|         2678313804|        0kbps|      7197|i LOVE @moon_devo...|      135|
|    6|1309515938773958656|    0nlyHoops|      3978|NBA Content Creat...|        1|
+-----+------------------+-------------+----------+--------------------+---------+
only showing top 20 rows
```

In [62]:
```
inf.count()
```

Out[62]: 408904

## Most Prolific Twitterers

In [53]:
```
original.count()
```

Out[53]: 11841802

In [56]:
```
retweet.count()
```

Out[56]: 17287060

In [192…
```
prolific = inf.select('user_name', 'user_descrip', 'followerct','count').sort(col('count').desc()).limit(50)
prolific.show()
```

```
[Stage 368:>                                                        (0 + 2) / 2]

+--------------+--------------------+----------+-----+
|     user_name|        user_descrip|followerct|count|
```

```
+--------------+--------------------+----------+-----+
|      sport9920|Welcom TV listing...|        94|19712|
|    ana92479235|Welcom TV listing...|       139|19679|
|       AgiwaraS| Enjoy your watching|        43|12649|
| AndrianyRahmah| hs game update news|       218|11725|
|  DennisStemmle|Founder - College...|      3804| 9236|
|   aljunasghost|i love sports hig...|        66| 8319|
|cintia238276122|both live and on ...|        69| 6683|
|SportsC62770366|We present online...|        73| 6441|
| robson89148161|  HIgh School Sports|        30| 6433|
|astro090_plo182|  high school sports|        15| 6134|
|     mbahucup84|Watch live sporti...|        31| 6075|
|  achue10983742|High School Footb...|        19| 5950|
|        Mesiga0|Sports Live Broad...|        35| 5846|
|   studyinnaija|Find All Private ...|        42| 5645|
|     1991Ananta|The most complete...|        25| 5556|
|       group_kq|THIS WEBSITE PROV...|       199| 5517|
|  cantiquesashy|i'm not perfect b...|        38| 5506|
|      lndrawari|High School Socce...|      1073| 5343|
|  freya23467530|Welcom TV listing...|        60| 5110|
|    TvSports112|               Sport|        59| 5054|
+--------------+--------------------+----------+-----+
only showing top 20 rows
```

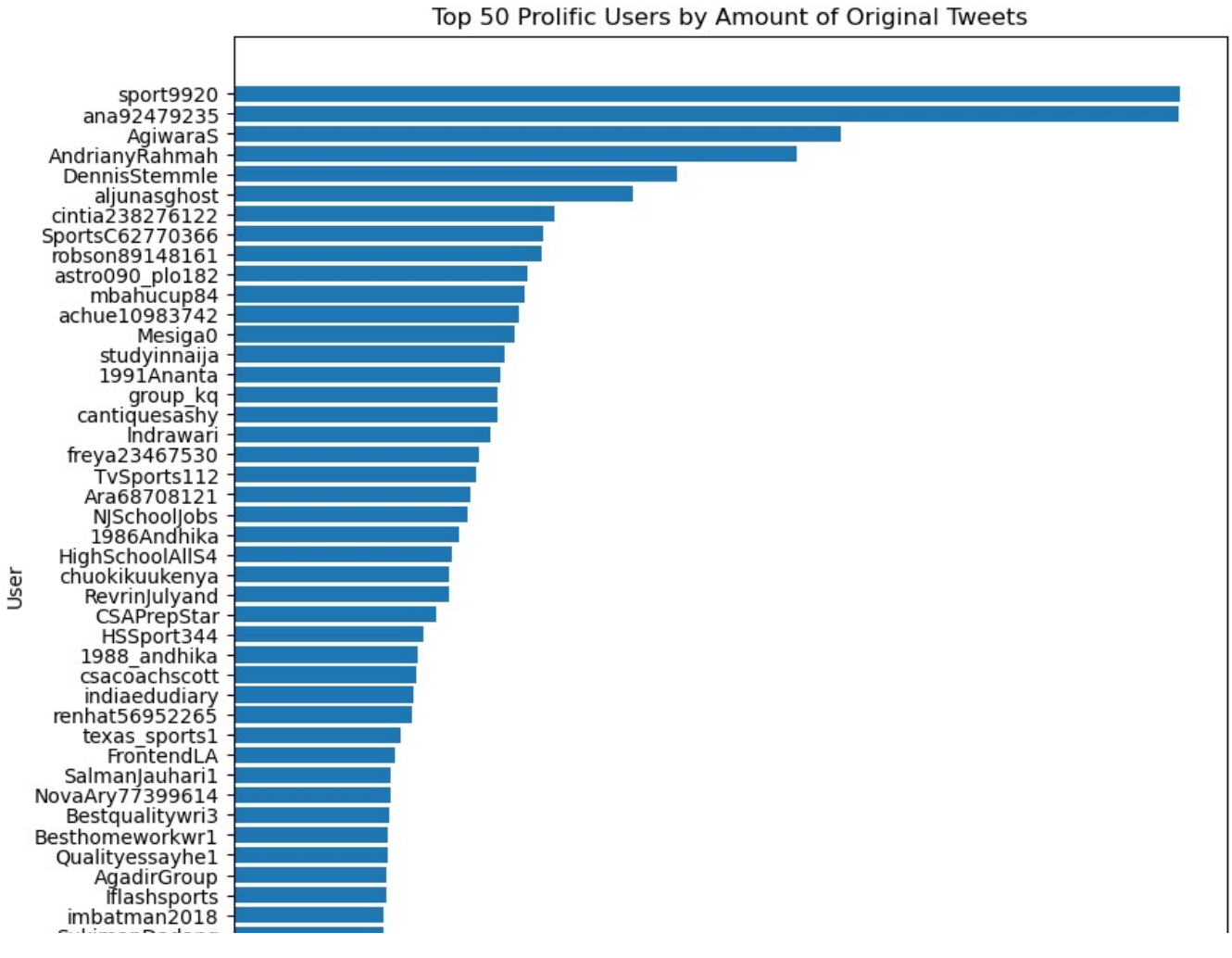
In [ ]:

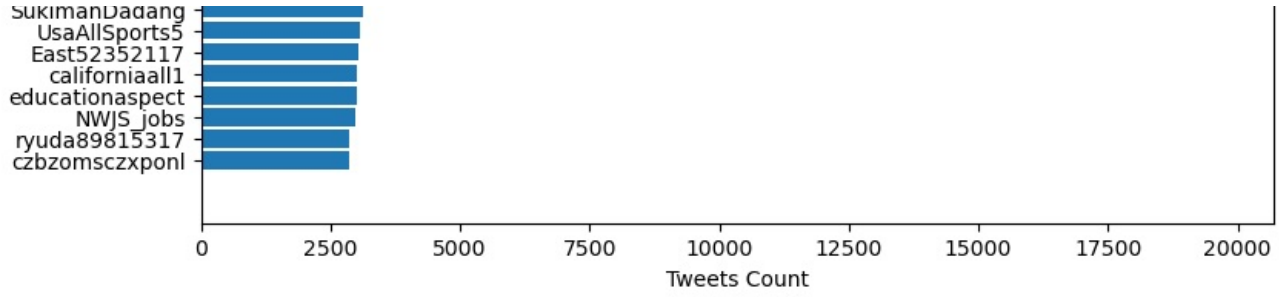
```
### Most of the top prolific twitterers are about sport
```

In [195…

```
p = prolific.toPandas()
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(p['user_name'], p['count'])
ax.invert_yaxis()
plt.xlabel('Tweets Count')
plt.ylabel('User')
plt.title('Top 50 Prolific Users by Amount of Original Tweets')
```

Out[195… Text(0.5, 1.0, 'Top 50 Prolific Users by Amount of Original Tweets')






## Most Retweet Twitterers

In [113…

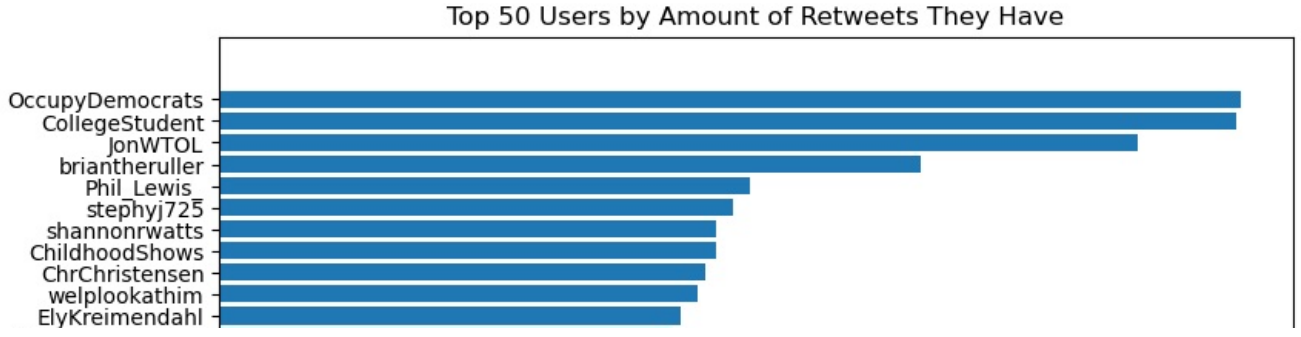
```
inf.select('user_name', 'user_descrip', 'followerct','total_rct').sort(col('total_rct').desc()).show()
```

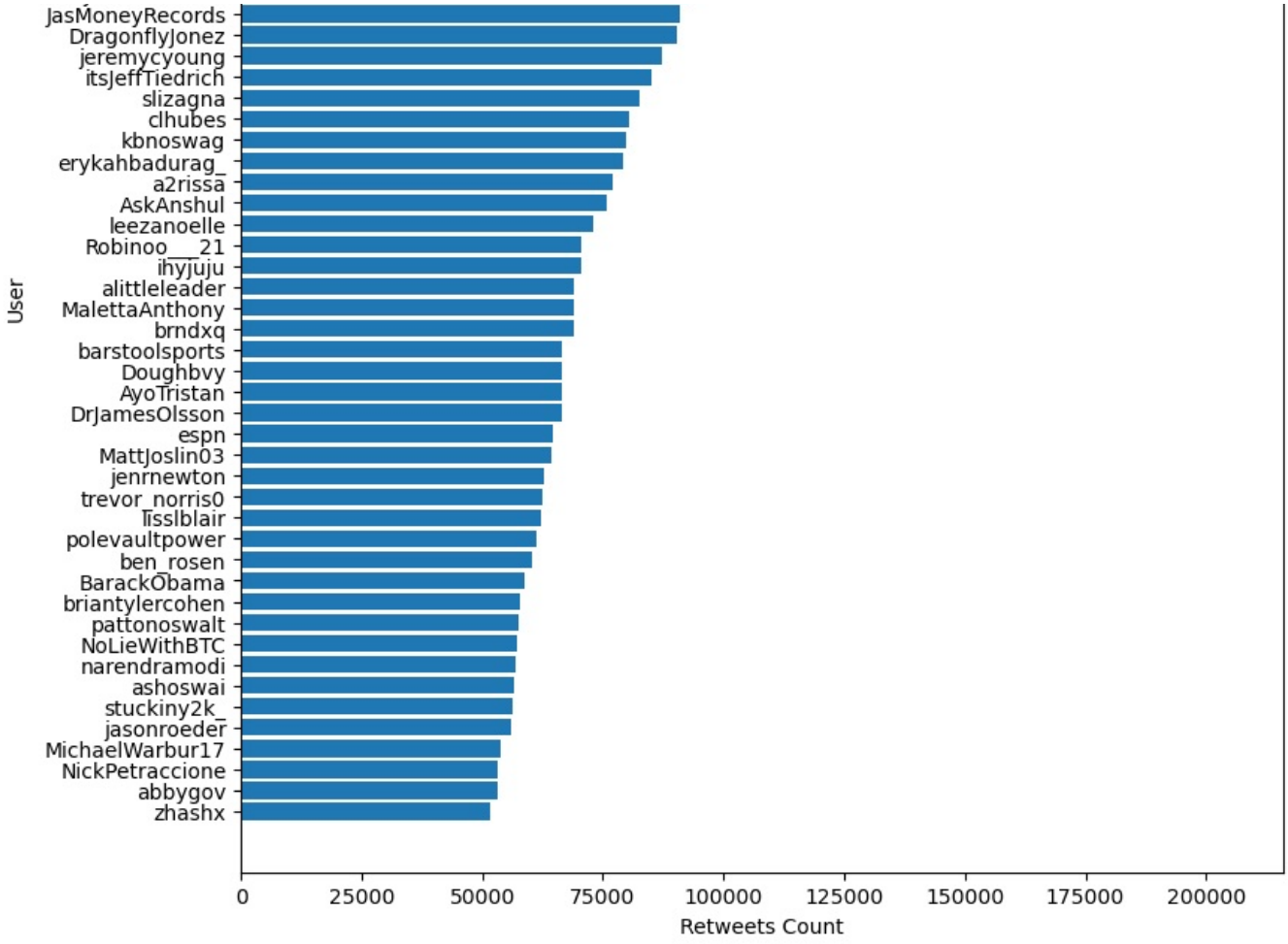
```
[Stage 254:>                                                        (0 + 2) / 2]
+---------------+--------------------+----------+---------+
|      user_name|        user_descrip|followerct|total_rct|
+---------------+--------------------+----------+---------+
|OccupyDemocrats|Pro-Democrat poli...|    502058|   205614|
| ColIegeStudent|Contact: CollegeS...|   1876945|   204652|
|        JonWTOL|NW Ohio native. R...|      2951|   184843|
| briantheruller|       turn notis on|    291035|   141139|
|    Phil_Lewis_|detroit native. s...|    237848|   106837|
|     stephyj725|exist loudly. ins...|      1389|   103432|
|  shannonrwatts|Founder of @MomsD...|    602813|    99920|
| ChildhoodShows|Reliving all of t...|    214893|    99920|
| ChrChristensen|American. Prof. o...|     34792|    97768|
|  welplookathim|Reluctant know-it...|      1060|    96276|
| ElyKreimendahl|writer, comic, qu...|    147944|    92736|
|JasMoneyRecords|Writer/Critic | L...|     16470|    90849|
| DragonflyJonez|You gonna just de...|    222012|    90184|
|   jeremycyoung|Senior Manager of...|     28855|    87301|
|itsJeffTiedrich|don't blame me, I...|   1001373|    85181|
|       slizagna|satisfaction not ...|      6849|    82663|
|        clhubes|editor @mcsweeney...|     38150|    80301|
|       kbnoswag|  @anus @barstoolyak|    271413|    79885|
| erykahbadurag_|Black supremacist...|       810|    79276|
|        a2rissa|avid wallows fan,...|      1888|    76952|
+---------------+--------------------+----------+---------+
only showing top 20 rows
```

In [199…

```
ct = inf.select('user_name', 'user_descrip', 'followerct','total_rct').sort(col('total_rct').desc()).limit(50)
c = ct.toPandas()
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(c['user_name'], c['total_rct'])
ax.invert_yaxis()
plt.xlabel('Retweets Count')
plt.ylabel('User')
plt.title('Top 50 Users by Amount of Retweets They Have')
```

Out[199… `Text(0.5, 1.0, 'Top 50 Users by Amount of Retweets They Have')`






In [198…

```
ct.select('user_name','user_descrip').collect()
```

Out[198…

```
[Row(user_name='OccupyDemocrats', user_descrip='Pro-Democrat political organization & news website. NY Times repo
rted that our reach dominated Trump on Facebook before his ban. Founder: @OmarRiverosays'),
 Row(user_name='ColIegeStudent', user_descrip='Contact: CollegeStudentofficial@gmail.com'),
 Row(user_name='JonWTOL', user_descrip='NW Ohio native. Reporter/Photographer for @WTOL11Toledo and @Go_419'),
 Row(user_name='briantheruller', user_descrip='turn notis on'),
 Row(user_name='Phil_Lewis_', user_descrip='detroit native. senior front page editor @huffpost.'),
 Row(user_name='stephyj725', user_descrip='exist loudly. insta: stephstagram.s'),
 Row(user_name='ChildhoodShows', user_descrip='Reliving all of the childhood shows and movies from late 90s and 2
000s.\n\nAlso tweets about childhood stars!'),
 Row(user_name='shannonrwatts', user_descrip='Founder of @MomsDemand, grassroots army of @Everytown / Author of F
ight Like a Mother / Boards @WeAdvancePeace, @EmergeAmerica / #MizzouMade'),
 Row(user_name='ChrChristensen', user_descrip='American. Prof. of Journalism at Stockholm Univ. Ph.D. Univ. of Te
xas at Austin. Contributor @TheLocalSweden. Editor Popular Communication (@comm_pop). 49ers.'),
 Row(user_name='welplookathim', user_descrip='Reluctant know-it-all, that knows everything except how fake feels.
Writer of Intriguingly Boring & Awesomely Imperfect. Thee OG #DadBod'),
 Row(user_name='ElyKreimendahl', user_descrip='writer, comic, queer. i live tweeted the birth of my baby. @funnyo
rdie @newrootsartists @humordarling mgmt: waldorf entertainment'),
 Row(user_name='JasMoneyRecords', user_descrip='Writer/Critic | Lots of work, lots of places | sandersjasmine1@gm
ail.com'),
 Row(user_name='DragonflyJonez', user_descrip="You gonna just decide that I'm drunk? Who are you to decide that I
'm drunk? I make that decision. That decision is between me and God."),
 Row(user_name='jeremycyoung', user_descrip="Senior Manager of Free Expression and Education @PENamerica. Histori
an. Free speech, academic freedom, children's rights, music, baseball. He/him. Views my own."),
 Row(user_name='itsJeffTiedrich', user_descrip="don't blame me, I voted for the email lady"),
 Row(user_name='slizagna', user_descrip='satisfaction not guaranteed. skeeball champion. earth worm rescuer. alt:
@slizbutsecret mutuals only, $20 for strangers. Venmo: slizagna she/her #BLM'),
 Row(user_name='clhubes', user_descrip='editor @mcsweeneys work in: @tnyshouts, @MADmagazine, @huffingtonpost, @b
ust_magazine, @themoth, @greatist, @runnersworld, @the_belladonnas Black Lives Matter'),
 Row(user_name='kbnoswag', user_descrip='@anus @barstoolyak'),
 Row(user_name='erykahbadurag_', user_descrip='Black supremacist & internet troll'),
 Row(user_name='a2rissa', user_descrip='avid wallows fan, lover of my friends'),
 Row(user_name='AskAnshul', user_descrip='| News Junkie | Politics | Foreign Affairs | National Security | Observ
er & Analyst | i tweet informative facts and opinions | http://www.youtube.com/AskAnshul'),
 Row(user_name='leezanoelle', user_descrip='rip grandpa wish you could hear nudyland'),
 Row(user_name='Robinoo___21', user_descrip="KSU 22'"),
 Row(user_name='ihyjuju', user_descrip='just a mf tweeting'),
```

```
 Row(user_name='alittleleader', user_descrip='Basically what would happen if a Powerpuff Girl became a Public Def
ender | trial attorney | she/her | must love sloths'),
 Row(user_name='MalettaAnthony', user_descrip='that guy'),
 Row(user_name='brndxq', user_descrip='19 | annoying nerd'),
 Row(user_name='barstoolsports', user_descrip='Viva La Stool - Order @roughnrowdy below for 20 fights + ring girl
contest Friday night at 8 pm ET'),
 Row(user_name='Doughbvy', user_descrip='The rose that grew from the concrete.'),
 Row(user_name='AyoTristan', user_descrip='In Alphabetical Order: Artist/Black/Christian/Dope/Funny/Gemini/INFJ/M
emes/Music/Politics/Ravenclaw/Sports/Tetris Master/Uncle/Writer/Yankee Hater/Zombie Hunter'),
 Row(user_name='DrJamesOlsson', user_descrip='Genetic Engineering, Johns Hopkins 2014,  Biomedical and Cancer Res
earch'),
 Row(user_name='espn', user_descrip='Serving sports fans. Anytime. Anywhere.'),
 Row(user_name='MattJoslin03', user_descrip='Busch Light Connoisseur with a thing for a smokin hot blonde. @Jewel
yBeff'),
 Row(user_name='jenrnewton', user_descrip='she/her, Teacher Educator, Assoc Professor, ungrader @teachingisintell
ectual on IG. Always curious, sometimes furious. Creatively non-compliant.'),
 Row(user_name='trevor_norris0', user_descrip='NOLA | Coast Life | Instagram : trevor_norris0 | Tik Tok - trevorn
orris'),
 Row(user_name='lisslblair', user_descrip='DE Social Studies Teacher, Adoptive Mom, Foodie, Gardener, Tulane & Vi
llanova History Alum, NBCT, 2021 DE Charter School TOY, 2019 Nat Geo GTF #teachSDGs'),
 Row(user_name='polevaultpower', user_descrip='Perpetually tilting at windmills. \n\nAlso @DefectiveBecca @Womens
Decathlon'),
 Row(user_name='ben_rosen', user_descrip='writer / reps: @echolakeent'),
 Row(user_name='BarackObama', user_descrip='Dad, husband, President, citizen.'),
 Row(user_name='briantylercohen', user_descrip='Political commentary. Over 1 billion views across YouTube, Facebo
ok, Instagram, Snapchat, TikTok. Host of @NoLieWithBTC podcast.'),
 Row(user_name='pattonoswalt', user_descrip='In 2022: @Netflix_Sandman, GASLIT, I LOVE MY DAD, new tour, album &
special. Podcast DID YOU GET MY TEXT? (link below). I like making stuff. IG: pattonoswalt'),
 Row(user_name='NoLieWithBTC', user_descrip='Podcast covering the top stories & interviews with the biggest names
in politics. Hosted by @briantylercohen'),
 Row(user_name='narendramodi', user_descrip='Prime Minister of India'),
 Row(user_name='ashoswai', user_descrip="Professor of Peace and Conflict Research\n@UU_Peace\n Uppsala University
\n'Right is Wrong' column, @gulf_news \nViews my own"),
 Row(user_name='stuckiny2k_', user_descrip='if peace is disturbed just block and watch industry on HBO'),
 Row(user_name='jasonroeder', user_descrip='Former senior writer/senior editor @theonion.'),
 Row(user_name='MichaelWarbur17', user_descrip='Actor, Cinephile, Docuphile, Drummer. Posting stuff I like - no p
andering. (backup acct - @TheMonologist)'),
 Row(user_name='NickPetraccione', user_descrip='Sports Anchor/Reporter for @abc27news. New York native. Syracuse
alum.'),
 Row(user_name='abbygov', user_descrip='she/her mgmt:kobrien@3arts.com personal:abby@abbygovindan.com'),
 Row(user_name='zhashx', user_descrip='hi hello')]
```

In [ ]:
```
#political orginizations, health organizations, educational workers, news outlets, social media influencers, othe
```

## Calculate Influence Score

In [ ]:
```
# the metrics I choose is to calculate infleunce score is to see the average number of retweet for each tweet,
# and how many followers they have
# retweet/total*follower_count
```

In [69]:
```
score = inf.withColumn('score', col('total_rct')/col('count')*col('followerct'))
```

In [111…
```
score.select('user_name', 'user_descrip','score').sort(col('score').desc()).show()
```

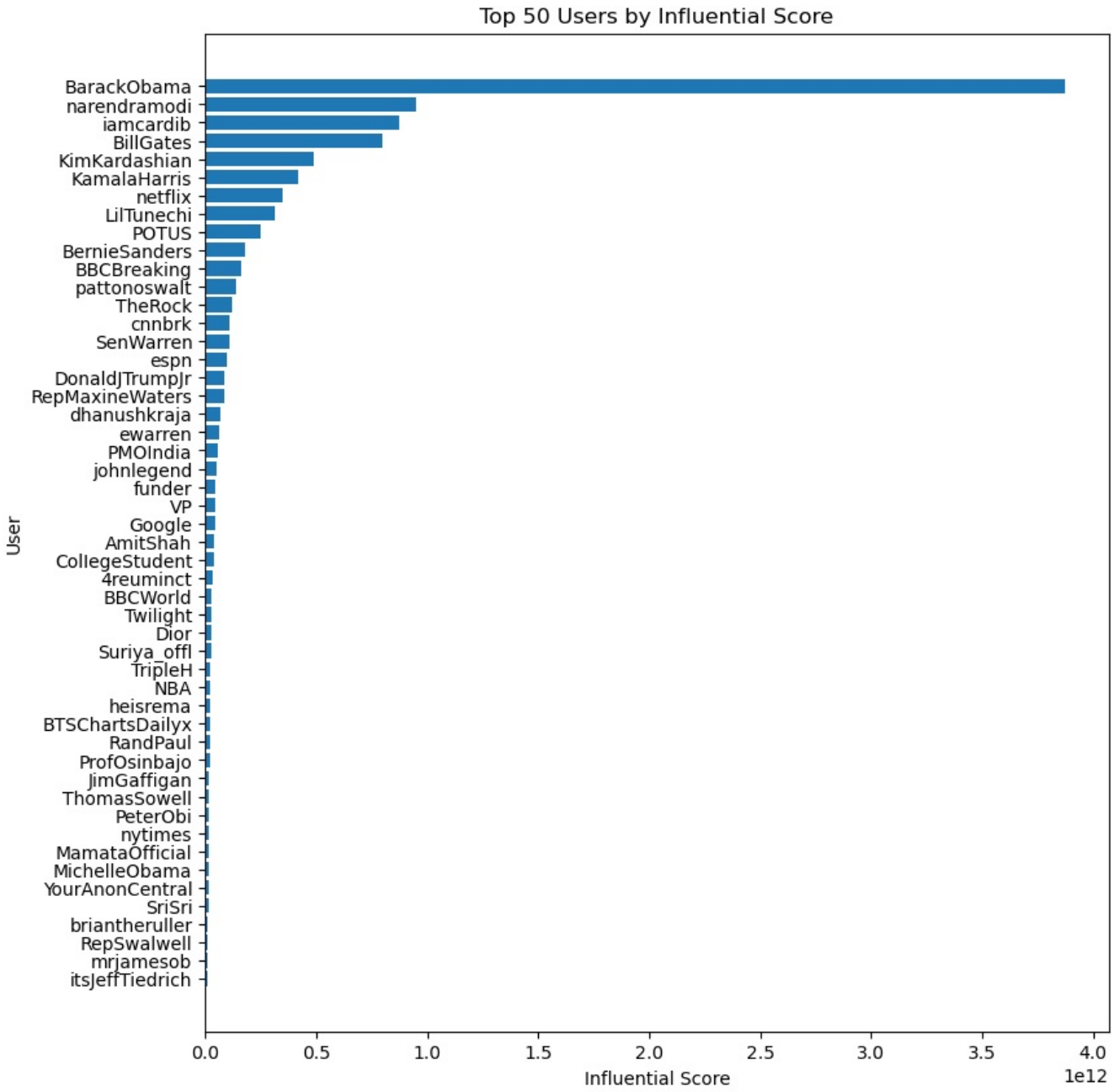
```
[Stage 252:>                                                        (0 + 2) / 2]
+---------------+--------------------+--------------------+
|      user_name|        user_descrip|               score|
+---------------+--------------------+--------------------+
|    BarackObama|Dad, husband, Pre...|  3.8754141030775E12|
|   narendramodi|Prime Minister of...|9.499095548982001E11|
|      iamcardib|       Mrs.DANGEROUS|     8.7274396714E11|
|      BillGates|Sharing things I'...|    7.97692446987E11|
|  KimKardashian|@SKKN @SKIMS @KAR...|    4.86270535764E11|
|   KamalaHarris|Fighting for the ...|     4.1755783108E11|
|        netflix|Naomi Watts. Jenn...|    3.48925762731E11|
|     LilTunechi|http://shoplilway...|    3.14771706176E11|
|          POTUS|46th President of...|2.472817499360000...|
|  BernieSanders|U.S. Senator for ...|1.794704084446666...|
|    BBCBreaking|Breaking news ale...|     1.5987526203E11|
|   pattonoswalt|In 2022: @Netflix...|    1.37490764138E11|
|        TheRock|             Founder|    1.23971070522E11|
|         cnnbrk|Breaking news fro...|    1.09996512884E11|
|      SenWarren|U.S. Senator, Mas...|    1.08244034736E11|
|           espn|Serving sports fa...|9.568168804541379E10|
| DonaldJTrumpJr|Future leader Min...|    8.89480753975E10|
|RepMaxineWaters|Proudly serving t...|     8.4208771368E10|
```

```
|   dhanushkraja|        ASURAN/Actor|     6.8177565632E10|
|        ewarren|U.S. Senator, for...|     6.0763314155E10|
+--------------+--------------------+--------------------+
only showing top 20 rows
```

In [10]:

```
sc = score.select('user_name', 'user_descrip','score').sort(col('score').desc()).limit(50)
s = sc.toPandas()
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(s['user_name'], s['score'])
ax.invert_yaxis()
plt.xlabel('Influential Score')
plt.ylabel('User')
plt.title('Top 50 Users by Influential Score')
```

Out[10]: Text(0.5, 1.0, 'Top 50 Users by Influential Score')




In [11]:

```
sc.select('user_name','user_descrip','score').collect()
```

Out[11]: [Row(user_name='BarackObama', user_descrip='Dad, husband, President, citizen.', score=3875414103077.5),
 Row(user_name='narendramodi', user_descrip='Prime Minister of India', score=949909554898.2001),
 Row(user_name='iamcardib', user_descrip='Mrs.DANGEROUS', score=872743967140.0),
 Row(user_name='BillGates', user_descrip="Sharing things I'm learning through my foundation work and other interests.", score=797692446987.0),
 Row(user_name='KimKardashian', user_descrip='@SKKN @SKIMS @KARDASHIANSHULU', score=486270535764.0),
 Row(user_name='KamalaHarris', user_descrip='Fighting for the people. Wife, Momala, Auntie. She/her. Official account is @VP.', score=417557831080.0),
 Row(user_name='netflix', user_descrip='Naomi Watts. Jennifer Coolidge. Mia Farrow. Noma Dumezweni. Margo Martindale.', score=348925762731.0),
 Row(user_name='LilTunechi', user_descrip='http://shoplilwayne.com', score=314771706176.0),
 Row(user_name='POTUS', user_descrip='46th President of the United States, husband to @FLOTUS, proud dad & pop. Tweets may be archived: http://whitehouse.gov/privacy\nText me: (302) 404-0880', score=247281749936.00003),
 Row(user_name='BernieSanders', user_descrip='U.S. Senator for Vermont. Not me, us.', score=179470408444.66666),
 Row(user_name='BBCBreaking', user_descrip='Breaking news alerts and updates from the BBC. For news, features, analysis follow @BBCWorld (international) or @BBCNews (UK). Latest sport news @BBCSport.', score=159875262030.0),
 Row(user_name='pattonoswalt', user_descrip='In 2022: @Netflix_Sandman, GASLIT, I LOVE MY DAD, new tour, album & special. Podcast DID YOU GET MY TEXT? (link below). I like making stuff. IG: pattonoswalt', score=137490764138.0),
 Row(user_name='TheRock', user_descrip='Founder', score=123971070522.0),
 Row(user_name='cnnbrk', user_descrip='Breaking news from CNN Digital. Now 64M strong. Check @cnn for all things CNN, breaking and more. Download the app for custom alerts: http://cnn.com/apps', score=109996512884.0),
 Row(user_name='SenWarren', user_descrip='U.S. Senator, Massachusetts. She/her/hers. Official Senate account.', score=108244034736.0),
 Row(user_name='espn', user_descrip='Serving sports fans. Anytime. Anywhere.', score=95681688045.41379),
 Row(user_name='DonaldJTrumpJr', user_descrip='Future leader Ministry of Truth, Father, Outdoorsman, Meme War General, founder: MxM News, Field Ethos Journal, Winning Team Publishing Pronouns:', score=88948075397.5),
 Row(user_name='RepMaxineWaters', user_descrip="Proudly serving the people of California's 43rd District in Congress. Chairwoman of the House Financial Services Committee (@FSCDems).", score=84208771368.0),
 Row(user_name='dhanushkraja', user_descrip='ASURAN/Actor', score=68177565632.0),
 Row(user_name='ewarren', user_descrip='U.S. Senator, former teacher. Wife, mom (Amelia, Alex, Bailey, @CFPB), grandmother, and Okie. She/her. Official campaign account.', score=60763314155.0),
 Row(user_name='PMOIndia', user_descrip='Office of the Prime Minister of India', score=60202026908.0),
 Row(user_name='johnlegend', user_descrip="Chrissy's husband. Father of Luna & Miles. #LoveInLasVegas The Residency in 2022. Tickets and VIP at http://johnlegend.com. Drink like me: @lve_wines", score=49243808406.0),
 Row(user_name='funder', user_descrip='Co-Founder & Executive Director @TheDemCoalition. @DworkinReport pod. Journalist. Investigator. Author. Musician. Obama alum. #TheResistance forever.', score=47450764575.0),
 Row(user_name='VP', user_descrip='Vice President of the United States. Wife to the first @SecondGentleman. Momala. Auntie. Fighting for the people.', score=47397636821.333336),
 Row(user_name='Google', user_descrip='#HeyGoogle', score=44577651456.0),
 Row(user_name='AmitShah', user_descrip='Union Home Minister, Minister of Cooperation and MP, Gandhinagar Lok Sabha. http://www.instagram.com/amitshahofficial', score=39403384720.0),
 Row(user_name='ColIegeStudent', user_descrip='Contact: CollegeStudentofficial@gmail.com', score=38412054814.0),
 Row(user_name='4reuminct', user_descrip='To be one with the stars.', score=31987644864.0),
 Row(user_name='BBCWorld', user_descrip="News, features and analysis from the World's newsroom. Breaking news, follow @BBCBreaking. UK news, @BBCNews. Latest sports news @BBCSport", score=30601443900.0),
 Row(user_name='Twilight', user_descrip='olivia rodrigo loves me', score=30218694160.0),
 Row(user_name='Dior', user_descrip='"Women, with their intuitive instinct, understood that I dreamed not only of making them more beautiful, but happier too." Christian Dior', score=29335059176.5),
 Row(user_name='Suriya_offl', user_descrip='Actor/Producer', score=28232960288.0),
 Row(user_name='TripleH', user_descrip='14-Time World Champion.  Chief Content Officer @WWE.', score=25118796202.0),
 Row(user_name='NBA', user_descrip='The 2022 #NBADraft presented by State Farm - Thursday, June 23 at 8:00pm/et on ABC and ESPN', score=24475558624.0),
 Row(user_name='heisrema', user_descrip='remabookings@gmail.com', score=23478507068.0),
 Row(user_name='BTSChartsDailyx', user_descrip='Welcome to news, charts source about the Princes of Global Pop & 2x Grammy-nominated @BTS_twt | Fan account', score=20988900189.0),
 Row(user_name='RandPaul', user_descrip='U.S. Senator for Kentucky | I fight for the Constitution, individual liberty and the freedoms that make this country great.', score=20129961187.0),
 Row(user_name='ProfOsinbajo', user_descrip="Official handle for Professor Yemi Osinbajo, SAN, Nigeria's Vice President. Tweets by him are signed YO", score=19673675362.0),
 Row(user_name='JimGaffigan', user_descrip='Male Model. COMEDY MONSTER on Netflix. Touring with all new material on THE FUN TOUR. My first wife is @jeanniegaffigan.', score=19352090256.0),
 Row(user_name='ThomasSowell', user_descrip="I'm not Thomas Sowell, but I own all of his books and tweet quotes from them. All tweets are direct quotes - @tsowellquotes. Get his latest book below.", score=17384760320.0),
 Row(user_name='PeterObi', user_descrip='Former Governor, Anambra State. Tweets by him are signed -PO.', score=16331579355.142857),
 Row(user_name='nytimes', user_descrip='News tips? Share them here: http://nyti.ms/2FVHq9v', score=16089206837.482014),
 Row(user_name='MamataOfficial', user_descrip='The Official Twitter of Mamata Banerjee, founder Chairperson All India Trinamool Congress. Honourable Chief Minister, West Bengal.', score=15989220840.0),
 Row(user_name='MichelleObama', user_descrip='Girl from the South Side and former First Lady. Wife, mother, dog lover. Always hugger-in-chief. #IAmBecoming', score=14792738782.5),
 Row(user_name='YourAnonCentral', user_descrip='Exposing Human Rights abuses. Actions Not Nouns. #YACnews\n\nSupport us: https://donorbox.org/launching-yac-news-site', score=14660341788.75),
 Row(user_name='SriSri', user_descrip='Mission: To See a Smile on Every Face; One World Family (Vasudhaiva Kutumbakam)', score=14021112170.0),
 Row(user_name='briantheruller', user_descrip='turn notis on', score=13692129621.666668),
 Row(user_name='RepSwalwell', user_descrip='Husband | Dad to Nelson, Cricket & Hank | Congressman | @HouseJudiciary @HouseIntel @HomelandDems | social media policy: http://bit.ly/3gxdzVm | #EndGunViolence', score=13687153076.0),
 Row(user_name='mrjamesob', user_descrip='Radio: http://lbc.co.uk Podcast: http://shorturl.at/tyC37 Game: http://

```
mysteryhour.co.uk', score=12955117104.0),
 Row(user_name='itsJeffTiedrich', user_descrip="don't blame me, I voted for the email lady", score=12185421930.42
8572)]
```

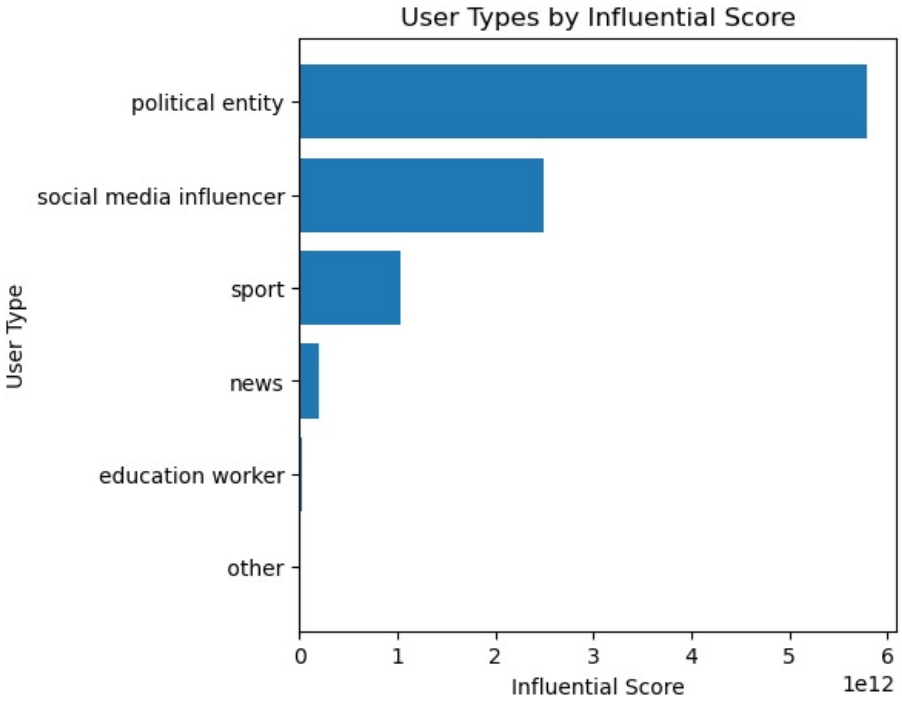
In [12]:
```
dic = {'user type':['political entity', 'social media influencer','sport', 'news', 'education worker', ' other'],
       'score':[5.8*(10**12),2.5*(10**12),10.3*(10**11),2*(10**11),3*(10**10), 1*(10**10)]}
df = pd.DataFrame.from_dict(dic)
df
```

Out[12]:

| | user type | score |
|---|---|---|
| 0 | political entity | 5.800000e+12 |
| 1 | social media influencer | 2.500000e+12 |
| 2 | sport | 1.030000e+12 |
| 3 | news | 2.000000e+11 |
| 4 | education worker | 3.000000e+10 |
| 5 | other | 1.000000e+10 |

In [15]:
```
fig, ax = plt.subplots(figsize =(5, 5))
ax.barh(df['user type'], df['score'])
ax.invert_yaxis()
plt.xlabel('Influential Score')
plt.ylabel('User Type')
plt.title('User Types by Influential Score')
```

Out[15]:
```
Text(0.5, 1.0, 'User Types by Influential Score')
```




## Topic Author Identification

In [1]:
```
import pandas as pd
import matplotlib.pyplot as plt
from pyspark.sql.functions import *
text = spark.read.parquet('gs://'+'msca-bdp-students-bucket/shared_data/chenfeng/project/key_eda/')
original = text.filter(col('retweet').isNull())
```

In [2]:
```
from pyspark.sql import functions as F
```

```
from pyspark.sql import types as t
from pyspark.sql.types import ArrayType, IntegerType, BooleanType

eng_ord=F.udf(lambda x: [ord(a) for a in x],t.ArrayType(IntegerType()))
def eng_filter(x):
    for index in range(len(x)):
        if x[index] > 128:
            return False
    return True

filter_udf = F.udf(eng_filter, BooleanType())
eng_ori = original.filter(filter_udf(eng_ord('full_text')) == True)
```

In [19]:

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

d = eng_ori.rdd.map(lambda x : x['full_text']).filter(lambda x: x is not None)
StopWords = stopwords.words("english")
# remove stop words
tokens = d\
    .map( lambda document: document.strip().lower())\
    .map( lambda document: re.sub("@[A-Za-z0-9_]+","", document))\
    .map( lambda document: re.sub(r'[^\w\s]', '', document))\
    .map( lambda document: re.split(" ", document))\
    .map( lambda word: [x for x in word if x not in StopWords])
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [5]:

```
d.count()
```

Out[5]: 7103950

In [6]:

```
eng_ori.count()
```

Out[6]: 7103950

In [20]:

```
t = tokens.zip(eng_ori.select('id').rdd.flatMap(lambda x:x))
df_tokens = t.toDF(['text_tokens','id'])
df_tokens.write.saveAsTable('d', mode = 'overwrite')
```

In [36]:

```
query = """
        select * from d where array_contains (text_tokens, 'k12')
        """
```

In [37]:

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

In [38]:

```
df.count()
```

Out[38]: 8563

In [41]:

```
df.limit(10).collect()
```

Out[41]:
```
[Row(text_tokens=['', 'prager', 'u', 'hillsdale', 'college', 'highly', 'regarded', 'k12', 'programs'], id=1511886
721268555778),
 Row(text_tokens=['likely', 'ensure', 'one', 'ever', 'learns', 'topics', 'teachers', 'k12', 'even', 'college', 'l
evel', 'lot', 'college', 'teachers', 'start', 'lower', 'education', 'teaching', 'technical', 'schools', 'httpstco
pqosbeow2n'], id=1512073363836964876),
 Row(text_tokens=['christian', 'college', 'building', 'nationwide', 'network', 'k12', 'public', 'charter', 'schoo
ls', 'thats', 'classical', 'american', 'httpstcomntbifvbwq'], id=1513927880576843779),
 Row(text_tokens=['\na', 'thread', 'recently', 'formed', 'sex', 'education', 'collaborative', '2020', 'made', '20
', 'groups', 'working', 'advance', 'comprehensive', 'sexuality', 'education', 'cse', 'k12', 'schools', 'nationwid
e', 'appear', 'want', 'cse', 'passed', 'federal', 'govt\nhttpstcoy0u7ipwb0w', 'httpstcoyyzbouljip'], id=151543408
6361407488),
 Row(text_tokens=['', 'also', 'lay', 'know', 'absolutely', 'nothing', 'schoolshow', 'k12', 'public', 'education',
'works'], id=1516265469052809216),
 Row(text_tokens=['', '', '', '', '', 'public', 'high', 'school', 'obama', 'elected', 'k12', 'public', 'schooler'
, 'ive', 'also', 'set', 'foot', 'public', 'school', 'multiple', 'times', 'since', '', 'one', 'parents', 'public',
'school', 'teacher', 'amp', 'kids', 'family', 'amp', 'friend', 'network', 'public', 'schoolers', 'sure', 'go'], i
d=1517202802711449600),
 Row(text_tokens=['university', 'mississippi', 'summer', 'programs', 'offer', 'fun', 'learning', 'k12', 'campers\
n\n', 'scholarship', 'opportunities', 'available', 'students', 'academic', 'levels\n\nfull', 'story', 'httpstco78
ayic06qn', 'httpstcowcswrtxkdt'], id=1518576829786861568),
 Row(text_tokens=['', '', 'equal', 'chance', 'succeed', 'life', 'get', 'opportunity', 'going', 'public', 'k12', '
school', 'free', 'want', 'extend', 'college', 'chose', 'college', 'could', 'afforda', 'major', 'would', 'get', 'j
ob', 'could', 'pay', 'loans'], id=1519459426926620672),
 Row(text_tokens=['', 'yeah', 'dont', 'pay', 'k12', 'teachers', 'millions', 'dollars', 'either', 'subsidize', 'co
llege'], id=1519830919917441025),
 Row(text_tokens=['', 'good', 'lower', 'immigration', 'fees', 'require', 'legal', 'immigrants', 'collect', 'benef
its', 'first', 'ten', 'years', 'citizenship', 'impose', 'tax', 'upon', 'fund', 'immigrant', 'k12', 'funding', 'co
llege', 'lending', 'well', 'funding', 'welfare', 'disabledelderly', 'immigrants'], id=1523083027995267072)]
```

In [42]:
```
k12 = df.join(original, df.id == original.id).select('user_id','user_name','user_descrip',df.id)
```

In [64]:
```
k12 = spark.read.parquet('gs://'+'msca-bdp-students-bucket/shared_data/chenfeng/project/k12/')
k12.show()
```

```
+------------------+--------------+--------------------+------------------+
|           user_id|     user_name|        user_descrip|                id|
+------------------+--------------+--------------------+------------------+
|1047877570740543489|        K12ssdb|Comprehensive dat...|1511412909838147584|
|         419966548|   HeidiHafeken|Advocate, adheren...|1511430882220990472|
|        3302663622|       NASAEPDC|The NASA STEM Edu...|1513879842944598031|
|1458122437917294600|   donna_calvey|I am not a doctor...|1515748426364997635|
|          78919708|  TouchstonesEd|Non-profit buildi...|1516066453493006337|
|        3487319294|       ssaisorg|#MeTooK12 creator...|1516515798742802433|
|        2793142971|   FlagstaffJon|Bernie supporter ...|1516825325480464387|
|1323654748357173248|  lightingnerd1|Yikes, uhh. Music...|1518343461782573058|
|        1938387535| RolfStraubhaar|Assistant Profess...|1518629414665342977|
|1234373598946590720|mortenson_nancy|Not my real name....|1520169427198570498|
| 829880699792551940|      MaEnraged|60's child, bohem...|1520531648013611010|
|1104023142043795457|      AmyrikalG|18 | multifandom ...|1524721145471250432|
|        2614569544|    ctachargers|K-12 public schoo...|1525083538269384706|
|        2874474721| previouslife17|Animal Lover ❤ □...|1527872587623194624|
|         223210328|          FMCSD|Fort McMurray Cat...|1529243368093630464|
|         342910848|     I_Am_Zackk|Free speech doesn...|1530624424168239104|
|        2432795983|    ISAachieves|Content & leaders...|1541490316695306241|
| 797268580387749888| klipschcollege|The Fred S. Klips...|1541509005226409984|
|1506948191090089990|OurSummerOfRage|Nationalism is an...|1543663545090560006|
|        3302663622|       NASAEPDC|The NASA STEM Edu...|1544320406114426886|
+------------------+--------------+--------------------+------------------+
only showing top 20 rows
```

In [74]:
```
k = k12.select('user_id').distinct()
kk = score.join(k, k.user_id == score.user_id,'inner').select(k.user_id, score.user_name, score.user_descrip,'sc
kk.count()
```

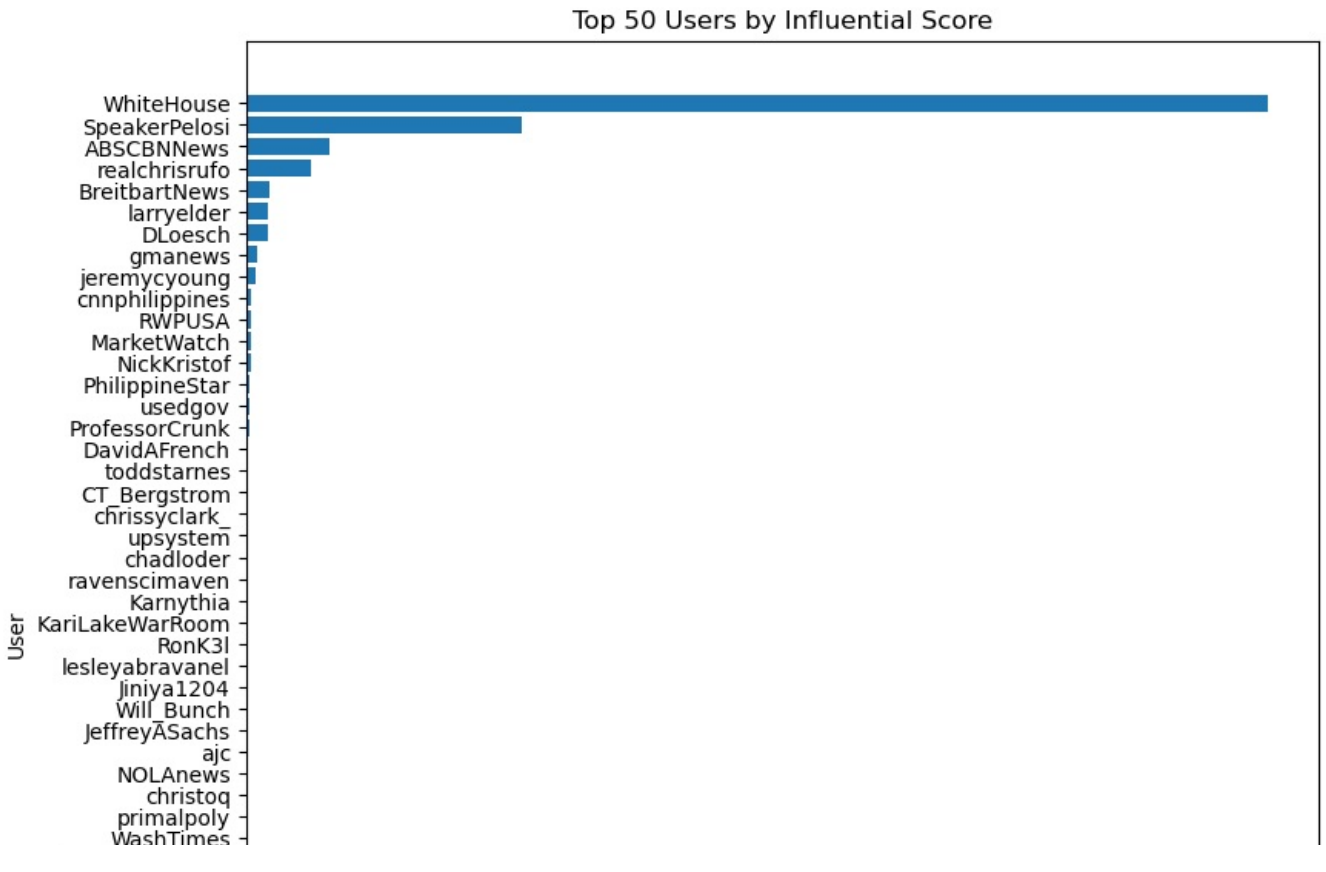
Out[74]: 2675

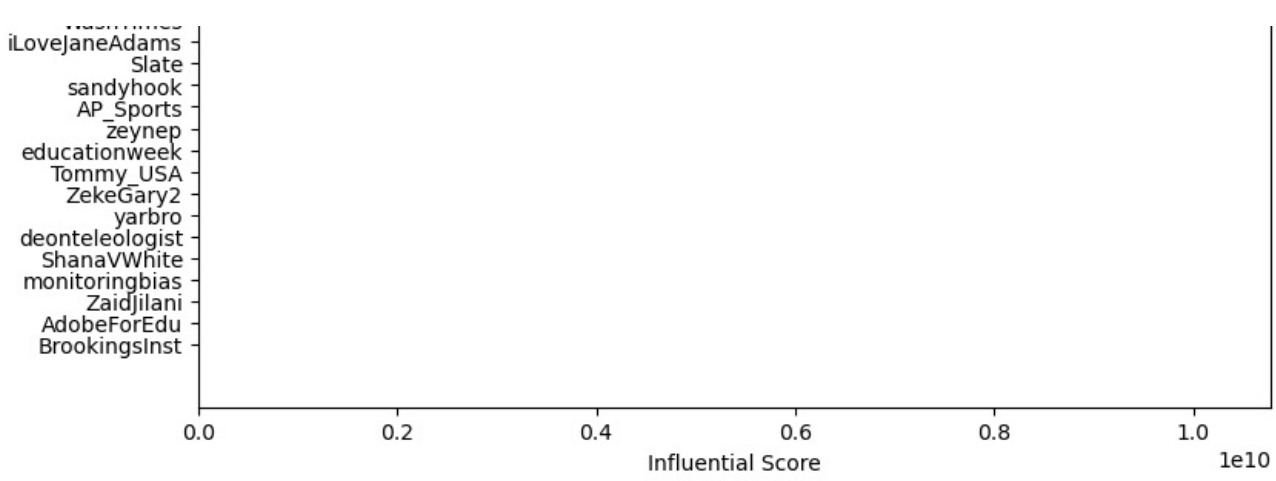
In [75]:
```
kk.sort(col('score').desc()).show()
```

```
[Stage 138:=============================================>           (4 + 1) / 5]
+------------------+-------------+--------------------+--------------------+
|           user_id|    user_name|        user_descrip|               score|
+------------------+-------------+--------------------+--------------------+
|1323730225067339784|   WhiteHouse|Welcome to the Bi...|1.026907747066666...|
|          15764644| SpeakerPelosi|Speaker of the Ho...|      2.7658038855E9|
|          15872418|   ABSCBNNews|Stories, video, a...| 8.361978978846154E8|
|        3066800573| realchrisrufo|Writer, filmmaker...| 6.504948614761904E8|
|         457984599| BreitbartNews|News, commentary,...|2.4026708525925925E8|
|         195271137|   larryelder|The Sage from Sou...|2.2747490133333334E8|
|           7702542|      DLoesch|Host of the natio...|       2.184114485E8|
|          39453212|      gmanews|Welcome to the of...|1.0962342109756097E8|
|         201826636|  jeremycyoung|Senior Manager of...| 9.688732134615384E7|
|        2811559122|cnnphilippines|News you can trus...| 5.467793930726257E7|
|        2863996955|       RWPUSA|Law Professor. Fo...| 5.004880184662577E7|
|            624413|   MarketWatch|News, personal fi...| 4.633792257142857E7|
|          17004618|   NickKristof|Oregon author, jo...|     4.29613716875E7|
|         472122299|PhilippineStar|The official Twit...|4.0037223751937985E7|
|          20437286|      usedgov|The official Unit...|      3.6134899875E7|
|         470708968|ProfessorCrunk|Professor. Contri...|2.7932862222222224E7|
|         240107748|  DavidAFrench|Senior Editor @Th...|           2.21837E7|
|          15515169|   toddstarnes|Follow the Todd S...|        2.20687302E7|
|        3238448948|  CT_Bergstrom|Biology professor...|2.1725433333333336E7|
|         263987952| chrissyclark_|words @dailycalle...|1.4718195176470589E7|
+------------------+-------------+--------------------+--------------------+
only showing top 20 rows
```

In [77]:
```
kk_df = kk.sort(col('score').desc()).limit(50).toPandas()
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(kk_df['user_name'], kk_df['score'])
ax.invert_yaxis()
plt.xlabel('Influential Score')
plt.ylabel('User')
plt.title('Top 50 Users by Influential Score')
```

Out[77]:
```
Text(0.5, 1.0, 'Top 50 Users by Influential Score')
```






In [79]:
```
kk.sort(col('score').desc()).select('user_name','user_descrip').limit(50).collect()
```

Out[79]:
```
[Row(user_name='WhiteHouse', user_descrip='Welcome to the Biden-Harris White House! Tweets may be archived: http://whitehouse.gov/privacy'),
 Row(user_name='SpeakerPelosi', user_descrip="Speaker of the House, focused on strengthening America's middle class and creating jobs; mother, grandmother, dark chocolate connoisseur."),
 Row(user_name='ABSCBNNews', user_descrip='Stories, video, and multimedia for Filipinos worldwide, from ABS-CBN News and Current Affairs'),
 Row(user_name='realchrisrufo', user_descrip='Writer, filmmaker, activist. Manhattan Institute and City Journal. Sign up for my newsletter: http://christopherrufo.com/newsletter.'),
 Row(user_name='BreitbartNews', user_descrip='News, commentary, and destruction of the political/media establishment.'),
 Row(user_name='larryelder', user_descrip="The Sage from South Central. Join me, because we've got a state and a country to save."),
 Row(user_name='DLoesch', user_descrip='Host of the nationally syndicated @DanaLoeschRadio #1 midday 12-3pET M-F. Bestselling author, 2A advocate, Daria, shadowbanned. http://danaloesch.substack.com'),
 Row(user_name='gmanews', user_descrip='Welcome to the official Twitter account of GMA News in the Philippines.'),
 Row(user_name='jeremycyoung', user_descrip="Senior Manager of Free Expression and Education @PENamerica. Historian. Free speech, academic freedom, children's rights, music, baseball. He/him. Views my own."),
 Row(user_name='cnnphilippines', user_descrip='News you can trust. @cnnphlife @sportsdeskph'),
 Row(user_name='RWPUSA', user_descrip='Law Professor. Former chief White House ethics lawyer 2005-07. Independent. Views are my own.'),
 Row(user_name='MarketWatch', user_descrip='News, personal finance & commentary from MarketWatch.\n\nFor customer support, visit https://customercenter.marketwatch.com'),
 Row(user_name='NickKristof', user_descrip='Oregon author, journalist and farmer, still sorting things out.'),
 Row(user_name='PhilippineStar', user_descrip="The official Twitter account of The Philippine STAR, flagship title of the country's most successful print media enterprise."),
 Row(user_name='usedgov', user_descrip='The official United States Department of Education Twitter account. Follows, Retweets and shared links ≠ endorsement.'),
 Row(user_name='ProfessorCrunk', user_descrip='Professor. Contributor @TheCut Books: Beyond Respectability| Eloquent Rage| The Crunk Feminist Collection| Feminist AF| Stand Up! (Aug. 2)'),
 Row(user_name='DavidAFrench', user_descrip='Senior Editor @TheDispatch, contributing writer @TheAtlantic, Co-host, Advisory Opinions and Good Faith podcasts, Iraq vet, married to @NancyAFrench.'),
 Row(user_name='toddstarnes', user_descrip='Follow the Todd Starnes Radio Show 12p-3p on GETTR: https://www.gettr.com/user/toddstarnes'),
 Row(user_name='CT_Bergstrom', user_descrip='Biology professor, @UW. Information flow in science and society. \n\nBook *Calling Bullshit*: http://tinyurl.com/fdcuvd7b\n\nI love ravens and crows. \n\nhe/him'),
 Row(user_name='chrissyclark_', user_descrip="words @dailycaller • @newsmax @iwn contributor • the underreported stories girl (opinions are my dad's bc I'm a conservative woman🙃) chrissy@dailycaller.com"),
 Row(user_name='upsystem', user_descrip='This is the official Twitter account of the University of the Philippines System, maintained by the Media and Public Relations Office.'),
 Row(user_name='chadloder', user_descrip='Writing about extremism, privacy, OSINT, and antifascism. (they/them)'),
 Row(user_name='ravenscimaven', user_descrip='Science communicator, molecular biologist, teaching science and building community 😀 AKA  shop  @smartypantsgear'),
 Row(user_name='Karnythia', user_descrip='Proud descendant of Hex Throwing Goons.  Amazons, Abolitionists and Activists &  NYT and Indie Bestseller Hood Feminism. Plays with fire. Rep @jillgrinberglit'),
 Row(user_name='KariLakeWarRoom', user_descrip='Official Campaign Twitter Account for the Trump-Endorsed Candidate for Arizona Governor. Text KARI to 70789. □□'),
 Row(user_name='RonK3l', user_descrip="Truth is stranger than fiction, but it is because Fiction is obliged to stick to possibilities; Truth isn't. MarkTwain\nAmerica First, ProLife, Jesus, Cartoonist"),
 Row(user_name='lesleyabravanel', user_descrip='NY-grown, FL-tanned, scribe, word nerd, TV junkie, game show champ, yenta, wife, twin mama, hot sauce collector, Bloody Mary maven &, says @NYPost, savvy gadfly'),
 Row(user_name='Jiniya1204', user_descrip='California ARMY born and raised in S. Korea □ 지니야:버터입덕 캘리아미\n□□□ Seokjin is my ultimate bias :) 석진아 많이 사랑한다! Fan account'),
 Row(user_name='Will_Bunch', user_descrip='National opinion columnist, Philadelphia Inquirer. New book AFTER THE IVORY TOWER FALLS (out now!) https://t.co/KaIAvvlnc5. Free weekly newsletter: https://t.co/8g2FAYQDEV'),
 Row(user_name='JeffreyASachs', user_descrip='Acadia University. Judicial politics, authoritarianism, Islam. Occasionally free speech on campus issues as well. Tweets my own -- and even then, just barely.'),
 Row(user_name='ajc', user_descrip='Our journalists can keep you informed with real, fact-based news because of s
```

```
ubscribers.\u202fLearn more: https://t.co/VnIn1RLW28 \n\nFor News tips and FAQs: https://t.co/syAWi1x4IM'),
 Row(user_name='NOLAnews', user_descrip='Latest news and updates from the Pulitzer Prize-winning newsroom of http
://NOLA.com & The Times-Picayune | New Orleans Advocate.'),
 Row(user_name='christoq', user_descrip='Followed by: @rosie @ElieNYC @PadmaLakshmi @RexChapman @MuellerSheWrote
@DavidCayJ @glennkirschner2 @ryangrim @Acyn @BenjaminPDixon #wtpBlue #Resist'),
 Row(user_name='primalpoly', user_descrip='Psych professor; wrote The Mating Mind, Spent, Mate, Virtue Signaling.
Themes: Evolution, sentience, civilization, EA, X risk, crypto. Wife: @sentientist.'),
 Row(user_name='WashTimes', user_descrip='Reliable reporting, hard-hitting analysis & breaking news, plus RTs fro
m our writers and reporters. Also at @WashTimesOpEd, @WashTimesLocal, @WashTimesSports'),
 Row(user_name='iLoveJaneAdams', user_descrip='Politics and Policy, Las Vegas | Former US House Candidate | Legis
lative Advocacy for Energy Independence, #Bitcoin, K-12 Financial Ed | Avocado Toast Lover'),
 Row(user_name='Slate', user_descrip='A daily magazine on the web. Subscribe to AMICUS, which has a special episo
de on the KBJ hearings on 3/19 for Slate Plus members only.'),
 Row(user_name='sandyhook', user_descrip='Take action and get involved to prevent school shootings and #EndGunVio
lence. Make the Promise to #ProtectOurKids. Account not monitored 24/7. RT ≠ endorsement.'),
 Row(user_name='AP_Sports', user_descrip='The top sports stories and insights from the global staff of The @AP, b
reaking news since 1846.\n\nPlease also follow @AP_Top25, @AP_NFL and @AP_Deportes!'),
 Row(user_name='zeynep', user_descrip='Complex systems, wicked problems. Society, technology, science and more. @
Columbia professor. @NYTimes columnist. My newsletter is @insight: https://t.co/6Ky01N9JwA'),
 Row(user_name='educationweek', user_descrip='Inspiring you through K-12 news, analysis, and opinion. Empowering
you to make a difference in your community.\n\nSign up for our newsletter: https://t.co/rFHMPLoZtc'),
 Row(user_name='Tommy_USA', user_descrip='Husband. Father. CEO, American Federation for Children @SchoolChoiceNow
. Fighting to give all kids K-12 #EducationFreedom #SchoolChoice'),
 Row(user_name='ZekeGary2', user_descrip='#MAGA, Save America, America 1st, Patriot, USAF Vet, NRA, Husband, Fath
er, Christian, #TRUMP2024'),
 Row(user_name='yarbro', user_descrip='TN State Senator - District 21 in Nashville. @BassBerrySims lawyer. Preds
fan, insomniac, husband of @tcyarbro, dad to Jack & Kate.'),
 Row(user_name='deonteleologist', user_descrip='PhD student (moral, social, political philosophy & epistemology (
esp. of race)) @CUNY_Philosophy | Research & Writing Fellow @AAPolicyForum | ⇨ #CRT'),
 Row(user_name='ShanaVWhite', user_descrip='My tweets are my own and usually are K12 education, computer science/
tech, and sports-related. “Try Jesus, not me.” -BW proverb'),
 Row(user_name='monitoringbias', user_descrip="Yup, it's me.\n\nData, culture, realism. Against PC, political tri
balism, and dogma. For free speech and cognitive decoupling. Saying the quiet part, loud."),
 Row(user_name='ZaidJilani', user_descrip="Solutions reporter at @NewsNation. Tweets reflect my opinions, not emp
loyers'.\nhttps://www.newsnationnow.com/author/zaid-jilani/"),
 Row(user_name='AdobeForEdu', user_descrip='To support, inspire and empower educators. Click below to check out o
ur Creative Challenges for students and educators! #AdobeEduCreative'),
 Row(user_name='BrookingsInst', user_descrip='Independent research and analysis on the most important policy issu
es in the world. Sign up for the daily Brookings Brief: https://brook.gs/3JuZxRE')]
```

In [ ]:

```
dic = {'user type':['political entity', 'social media influencer','sport', 'news', 'education worker', ' other'],
      'score':[5.8*(10**12),2.5*(10**12),10.3*(10**11),2*(10**11),3*(10**10), 1*(10**10)]}
df = pd.DataFrame.from_dict(dic)
df
fig, ax = plt.subplots(figsize =(5, 5))
ax.barh(df['user type'], df['score'])
ax.invert_yaxis()
plt.xlabel('Influential Score')
plt.ylabel('User Type')
plt.title('User Types by Influential Score')
```

In [ ]:

In [ ]:

In [ ]:

## Geographical Trends

In [115…

```
loc = text.groupby('user_loc').agg(count('user_loc').alias('loc_ct'))
loc.sort(col('loc_ct').desc()).show()
```

```
[Stage 257:==================================>                    (8 + 4) / 13]
```

```
+--------------+------+
|      user_loc|loc_ct|
+--------------+------+
|  United States|479269|
| Lagos, Nigeria|157291|
|       Nigeria|141868|
|  united states|135952|
|Los Angeles, CA|115640|
```

```
|          India|111718|
|            USA|104641|
|California, USA| 98784|
|     Texas, USA| 89596|
|    Chicago, IL| 86856|
|    Atlanta, GA| 86498|
|    Houston, TX| 86129|
|   Florida, USA| 77657|
| Nairobi, Kenya| 75089|
|        she/her| 74262|
| Washington, DC| 71899|
|London, England| 67606|
|         Canada| 62984|
|   New York, NY| 59052|
| United Kingdom| 57306|
+---------------+------+
only showing top 20 rows
```

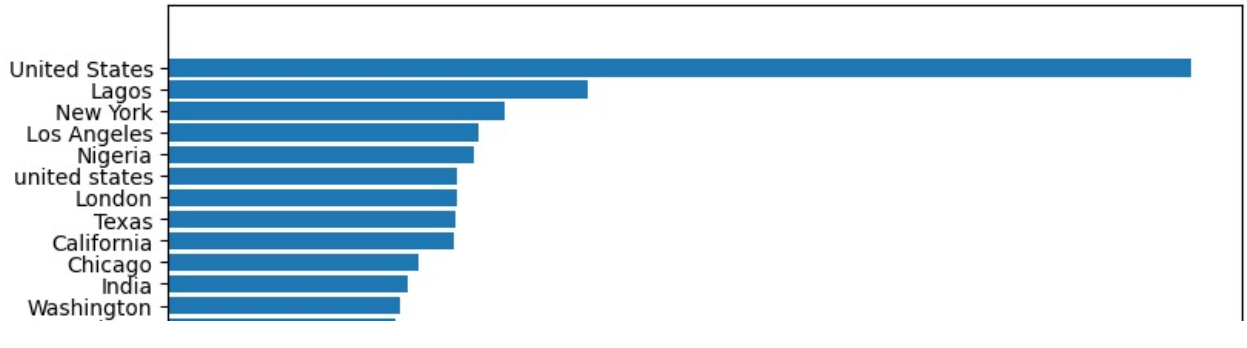
In [170…

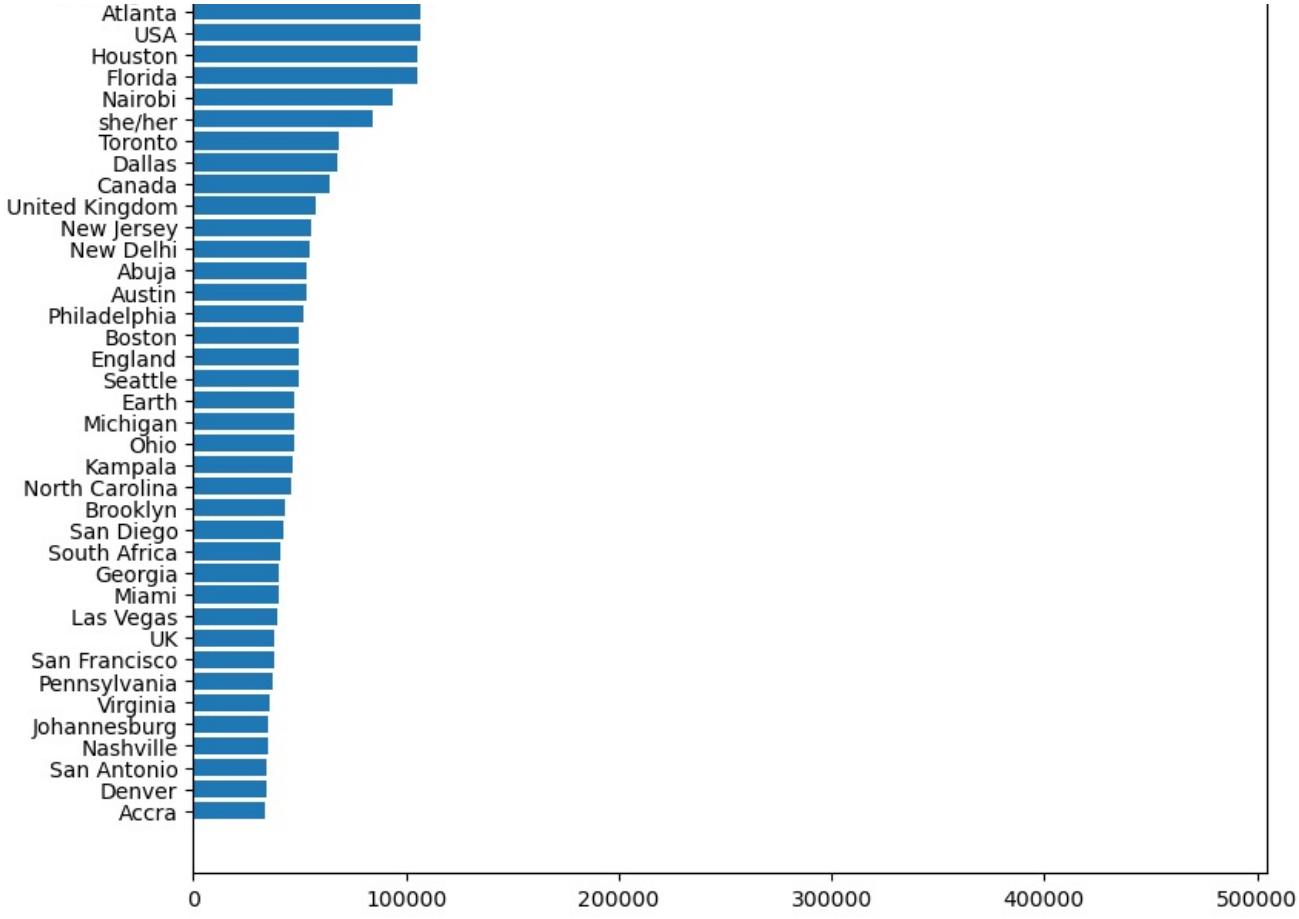
```
# only extract the top 50 locations
b = loc.select(substring_index(col('user_loc'),',', 1).alias('loc'),'loc_ct')
b = b.groupby('loc').agg(sum('loc_ct').alias('ct')).sort(col('ct').desc()).limit(50)
b.show()
```

```
[Stage 342:=======================================>                 (9 + 4) / 13]
+-------------+------+
|          loc|    ct|
+-------------+------+
|United States|480493|
|        Lagos|196959|
|     New York|158179|
|  Los Angeles|145666|
|      Nigeria|143681|
|united states|135996|
|       London|135714|
|        Texas|134830|
|   California|134039|
|      Chicago|117673|
|        India|112849|
|   Washington|109128|
|      Atlanta|107162|
|          USA|106970|
|      Houston|105650|
|      Florida|105026|
|      Nairobi| 93866|
|      she/her| 84459|
|      Toronto| 68293|
|       Dallas| 67996|
+-------------+------+
only showing top 20 rows
```

In [176…

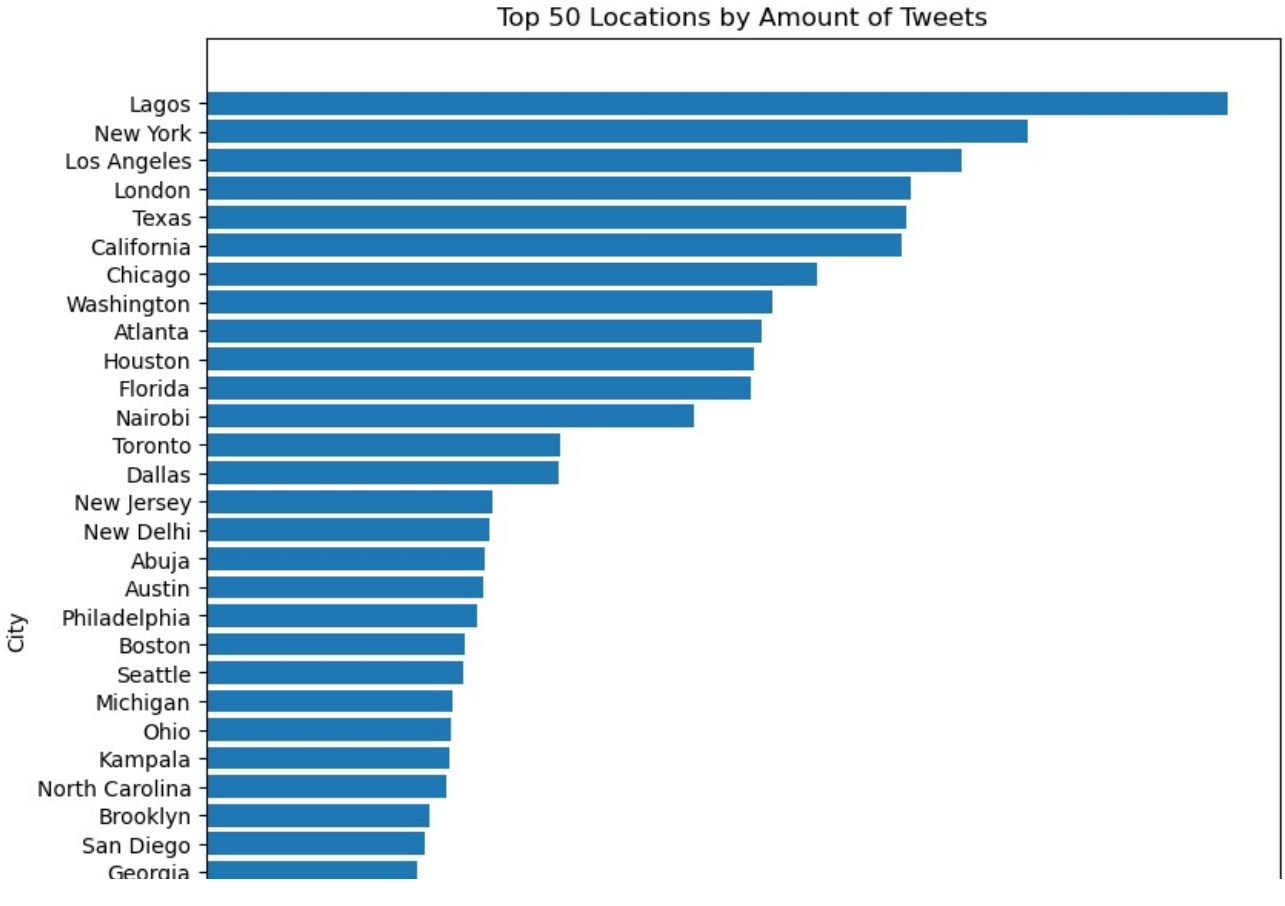
```
location = b.toPandas()
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(location['loc'],location['ct'])
ax.invert_yaxis()
plt.show()
```

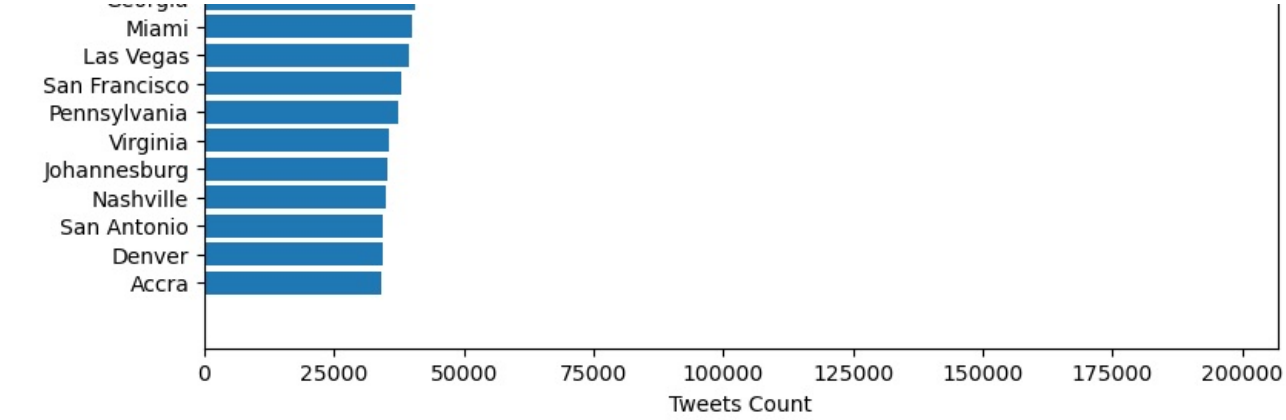





In [186…

```
# remove country
l = ['USA','United States','united states', 'UK','United Kingdom','India','England','South Africa',
     'Nigeria','Canada','she/her', 'Earth']
city = location[~location['loc'].isin(l)]
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(city['loc'],city['ct'])
ax.invert_yaxis()
plt.xlabel('Tweets Count')
plt.ylabel('City')
plt.title('Top 50 Locations by Amount of Tweets')
plt.show()
```





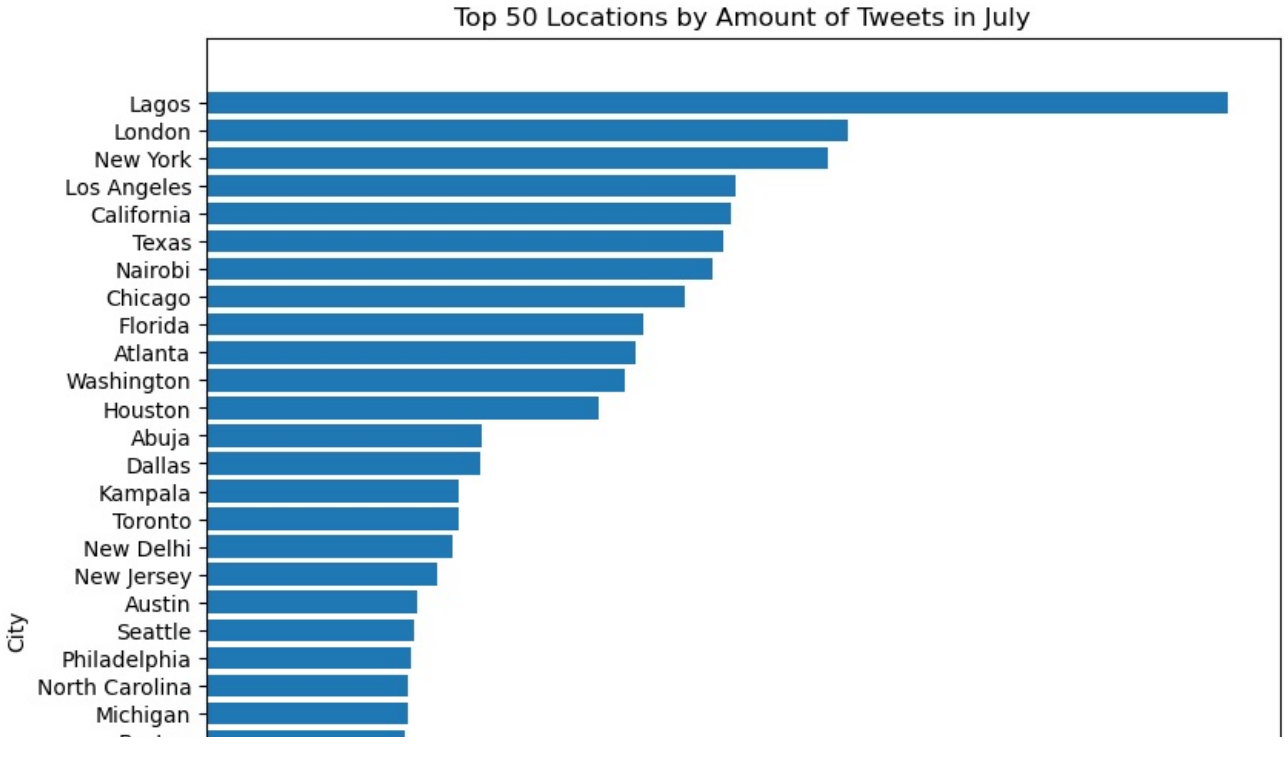

In [ ]:
```
# It has the same trend as population ranking with NY the most then LA, Chicago and Houston, Dellas, Miami
# Most of them are from main city, and they occupied the top list of this ranking
# Then the second tier are some cities closed by the main city, such as Michigan, Boston, San Diego
# Some of them are from UK, Africa(Nigeria, South Africa), India
```

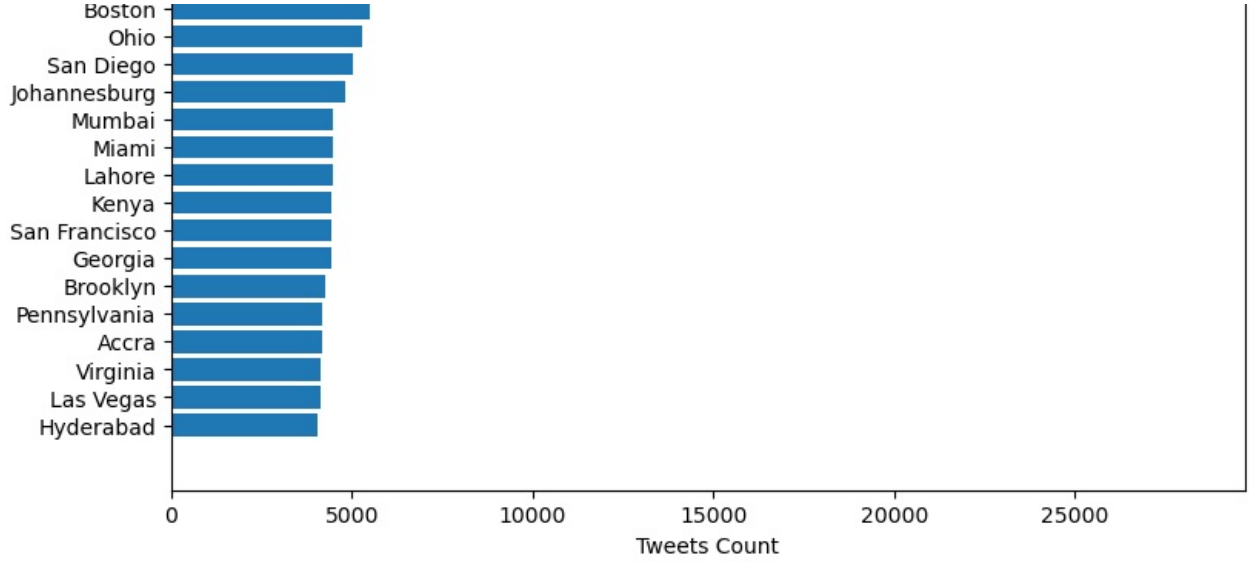
## Geographical Progression

In [2]:
```
t = text.withColumn('mnth', substring_index(col('created_at'),' ', 3)).drop('created_at')
t = t.withColumn('mnth', substring_index(col('mnth'),' ',-2))
t = t.withColumn('mnth', substring_index(col('mnth'),' ',1))
t = t.withColumn("mnth",from_unixtime(unix_timestamp(col("mnth"),'MMM'),'MM'))
```

In [3]:
```
july = t.filter(col('mnth')=='07')
aug = t.filter(col('mnth')=='08')
sep = t.filter(col('mnth')=='09')
```

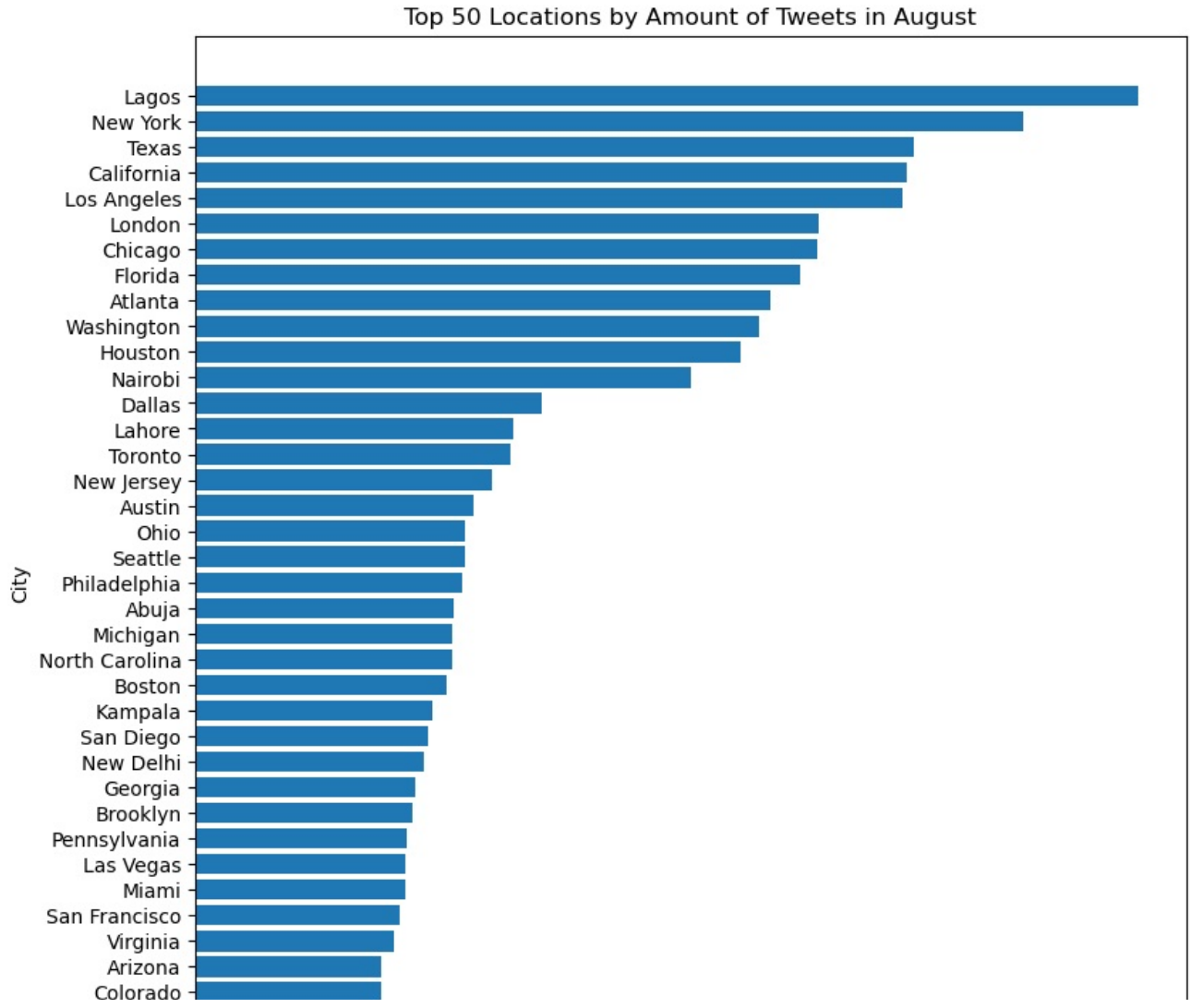
In [4]:
```
loc = july.groupby('user_loc').agg(count('user_loc').alias('loc_ct'))
b = loc.select(substring_index(col('user_loc'),',', 1).alias('loc'),'loc_ct')
b = b.groupby('loc').agg(sum('loc_ct').alias('ct')).sort(col('ct').desc()).limit(50)
location = b.toPandas()
l = ['USA','United States','united states', 'UK','United Kingdom','India','England','South Africa',
     'Nigeria','Canada','she/her', 'Earth']
city = location[~location['loc'].isin(l)]
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(city['loc'],city['ct'])
ax.invert_yaxis()
plt.xlabel('Tweets Count')
plt.ylabel('City')
plt.title('Top 50 Locations by Amount of Tweets in July')
plt.show()
```

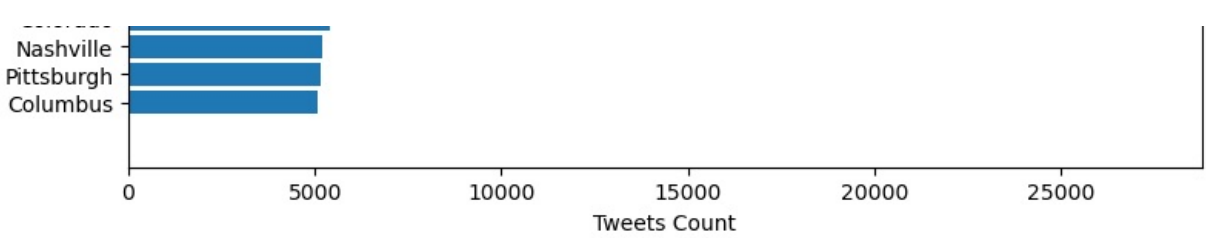





In [5]:

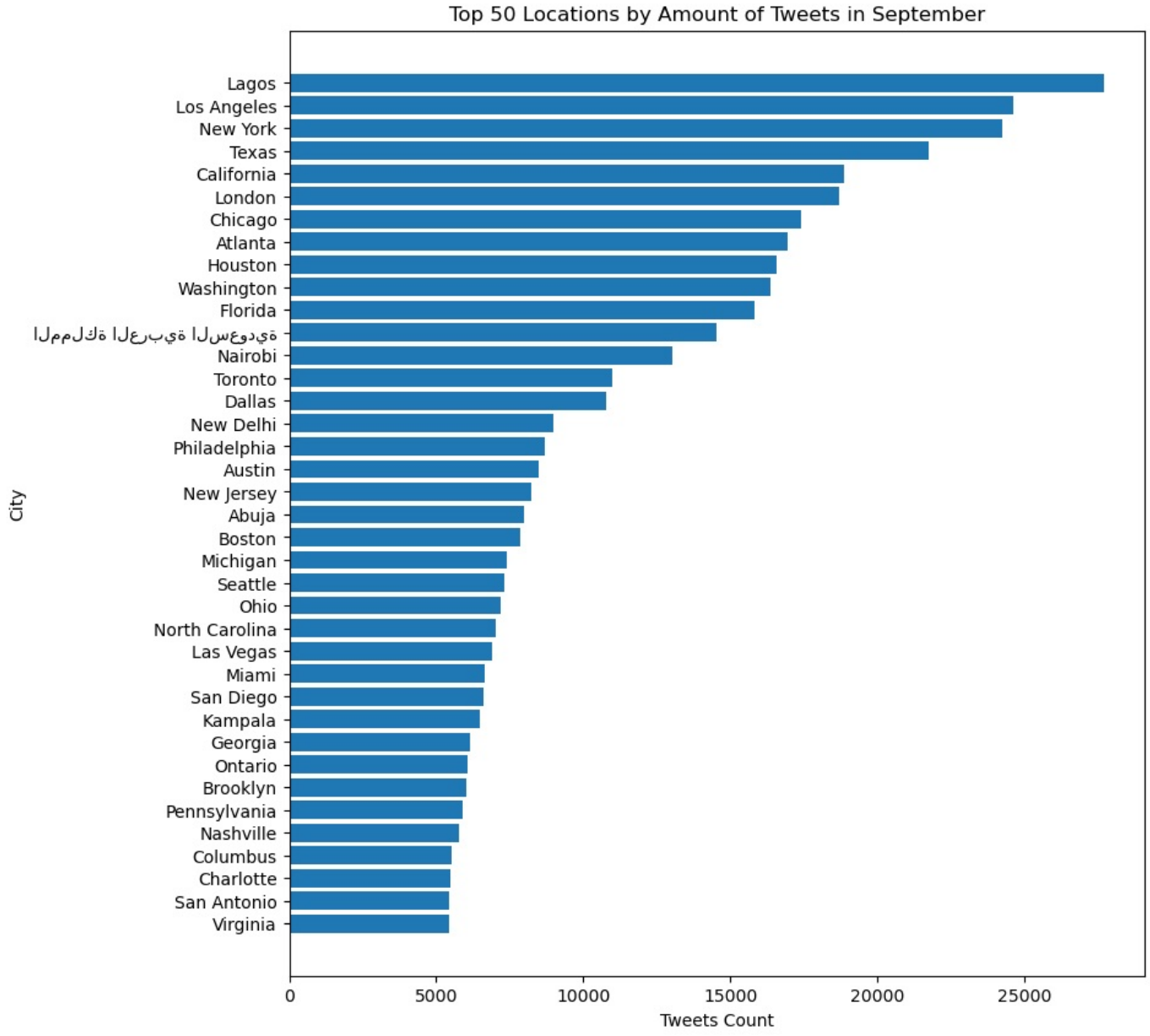
```
loc = aug.groupby('user_loc').agg(count('user_loc').alias('loc_ct'))
b = loc.select(substring_index(col('user_loc'),',', 1).alias('loc'),'loc_ct')
b = b.groupby('loc').agg(sum('loc_ct').alias('ct')).sort(col('ct').desc()).limit(50)
location = b.toPandas()
l = ['USA','United States','united states', 'UK','United Kingdom','India','England','South Africa',
     'Nigeria','Canada','she/her', 'Earth']
city = location[~location['loc'].isin(l)]
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(city['loc'],city['ct'])
ax.invert_yaxis()
plt.xlabel('Tweets Count')
plt.ylabel('City')
plt.title('Top 50 Locations by Amount of Tweets in August')
plt.show()
```






In [6]:

```
loc = sep.groupby('user_loc').agg(count('user_loc').alias('loc_ct'))
b = loc.select(substring_index(col('user_loc'),',', 1).alias('loc'),'loc_ct')
b = b.groupby('loc').agg(sum('loc_ct').alias('ct')).sort(col('ct').desc()).limit(50)
location = b.toPandas()
l = ['USA','United States','united states', 'UK','United Kingdom','India','England','South Africa',
     'Nigeria','Canada','she/her', 'Earth']
city = location[~location['loc'].isin(l)]
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(city['loc'],city['ct'])
ax.invert_yaxis()
plt.xlabel('Tweets Count')
plt.ylabel('City')
plt.title('Top 50 Locations by Amount of Tweets in September')
plt.show()
```

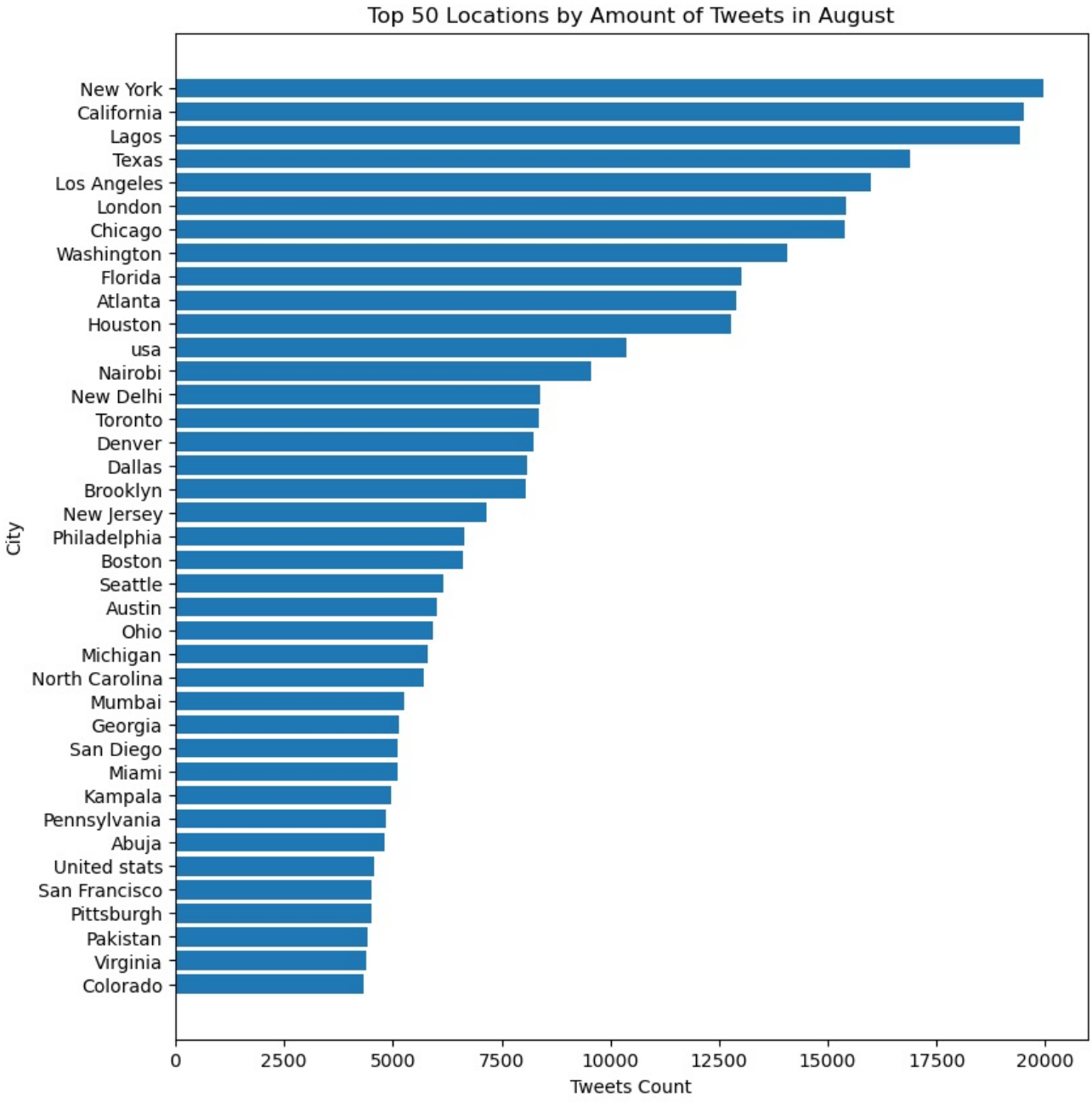



In [7]:

```
april = t.filter(col('mnth')=='04')
may = t.filter(col('mnth')=='05')
```

In [8]:

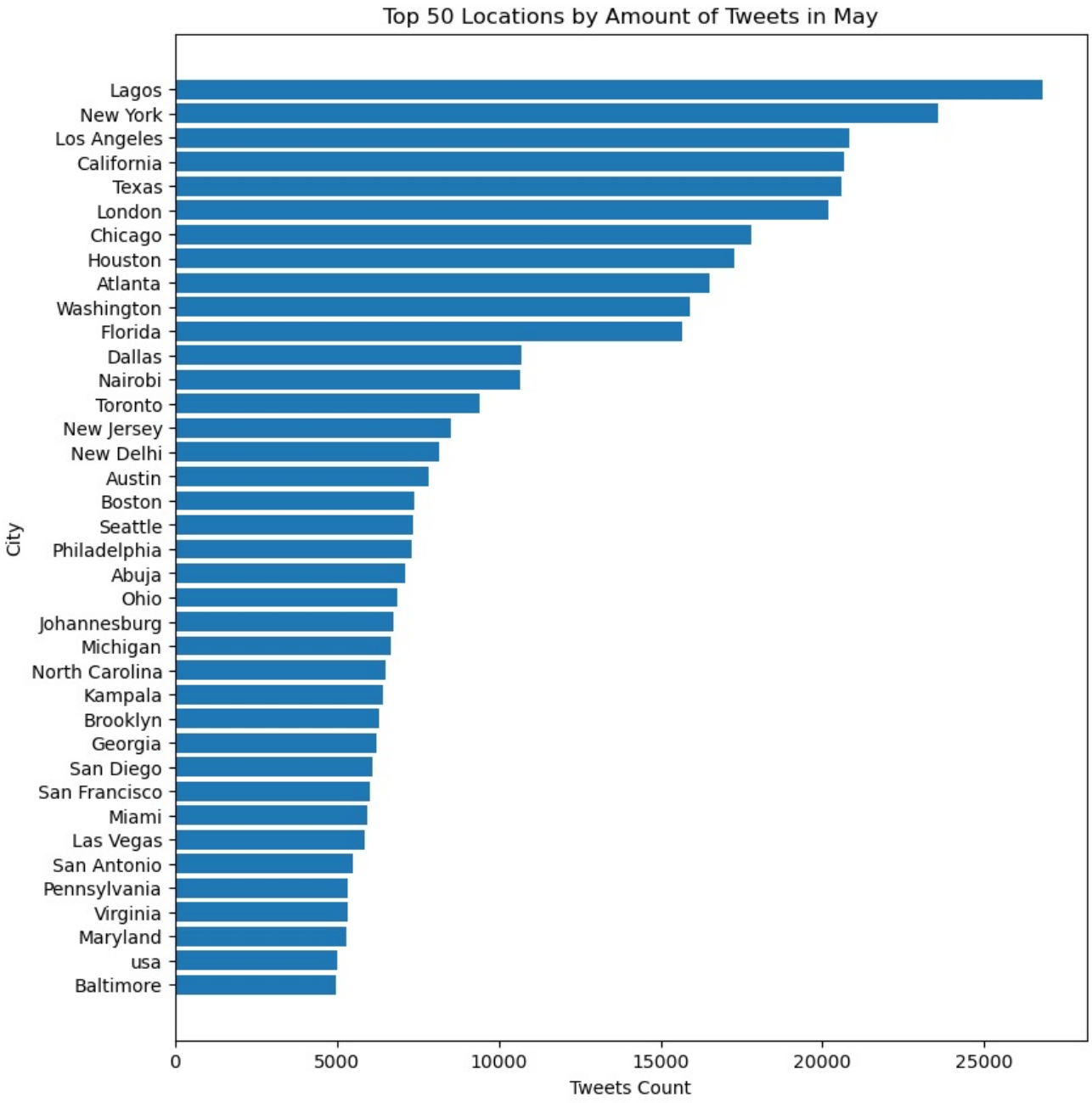
```
loc = april.groupby('user_loc').agg(count('user_loc').alias('loc_ct'))
b = loc.select(substring_index(col('user_loc'),',', 1).alias('loc'),'loc_ct')
b = b.groupby('loc').agg(sum('loc_ct').alias('ct')).sort(col('ct').desc()).limit(50)
location = b.toPandas()
l = ['USA','United States','united states', 'UK','United Kingdom','India','England','South Africa',
     'Nigeria','Canada','she/her', 'Earth']
city = location[~location['loc'].isin(l)]
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(city['loc'],city['ct'])
ax.invert_yaxis()
plt.xlabel('Tweets Count')
plt.ylabel('City')
plt.title('Top 50 Locations by Amount of Tweets in April')
plt.show()
```




In [9]:

```
loc = may.groupby('user_loc').agg(count('user_loc').alias('loc_ct'))
b = loc.select(substring_index(col('user_loc'),',', 1).alias('loc'),'loc_ct')
b = b.groupby('loc').agg(sum('loc_ct').alias('ct')).sort(col('ct').desc()).limit(50)
location = b.toPandas()
l = ['USA','United States','united states', 'UK','United Kingdom','India','England','South Africa',
     'Nigeria','Canada','she/her', 'Earth']
city = location[~location['loc'].isin(l)]
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(city['loc'],city['ct'])
ax.invert_yaxis()
plt.xlabel('Tweets Count')
plt.ylabel('City')
```

```
plt.title('Top 50 Locations by Amount of Tweets in May')
plt.show()
```




## Topic Progression

In [20]:
```
eng.write.saveAsTable('d', mode = 'overwrite')
```

In [68]:
```
query = """
        select * from d where array_contains (tokens, 'florida') \
        or array_contains (tokens, 'math') \
        or array_contains (tokens, 'book') \
        or array_contains (tokens, 'ban')
        """
```

In [69]:
```
df = spark.sql(query)
```

In [24]:
```
df.show()
```

```
+--------------------+--------------------+
```

```
|              tokens|                 id|
+--------------------+-------------------+
|[lakeland, christ...|1511458095557345281|
|[tyler, atkinson,...|1511583844067069953|
|[simmons, math, l...|1511666734641651713|
|[sunset, timelaps...|1511869310179885075|
|[contrast, first,...|1511942471378345989|
|[shocked, swore, ...|1512131393765220362|
|[aucilla, christi...|1512165961675476993|
|[name, coach, lan...|1512212986131365889|
|[writing, homewor...|1512268414915518464|
|[university, cent...|1512389358212505606|
|[oxford, universi...|1512416319467139077|
|[keiser, universi...|1512475550237003781|
|[bronson, yonge, ...|1512527047318388736|
|[palmer, trinity,...|1512527304328560642|
|[might, rabbi, ma...|1512542108061782023|
|[might, rabbi, ma...|1512542108061782023|
|[might, rabbi, ma...|1512542108061782023|
|[might, rabbi, ma...|1512542108061782023|
|[college, complet...|1512757581034373124|
|[college, checked...|1512820311753781249|
+--------------------+-------------------+
only showing top 20 rows
```

In [21]:
```
eng.count()
```

Out[21]: 7103950

In [70]:
```
df.count()
```

Out[70]: 109413

In [75]:
```
tt = df.join(text, df.id == text.id,'inner').select(text.id, text.tokens,'user_id','user_name','user_descrip','re
tt.show()
```

```
[Stage 227:=============================================================> (65 + 2) / 67]
+-------------------+--------------------+-------------------+--------------+--------------------+-------+--------------------+
|                 id|              tokens|            user_id|     user_name|        user_descrip|retweet|           full_text|
+-------------------+--------------------+-------------------+--------------+--------------------+-------+--------------------+
|1511334512759779328|[daniel, greenfie...|          279166471| SeraphicSecret|Emmy Award winnin...|   null|Daniel Greenfield...|
|1511454787530862594|[hood, niggas, wo...|1323452208692764673|        KINGCMo|Makin’ Treasure O...|   null|Hood Niggas would...|
|1511488396799234051|[florida, gulf, c...|         4164446764|     FGCUWxSTEM|Providing real-ti...|   null|Florida Gulf Coas...|
|1511686149424791559|[european, editio...|         1347250338|RichardGCorbett|Formerly: Yorks&H...|   null|The European Unio...|
|1512090926990204930|[great, piece, ed...|          201826636|   jeremycyoung|Senior Manager of...|   null|Great piece on ed...|
|1512478292149952520|[great, college, ...|           79228319|    StLHandyMan|I believe the pow...|   null|@WyrdGrl @hollyma...|
|1512492327188803585|[close, high, sch...|1494504264609284099|BiggusDickus444|GIFs n Memes make...|   null|@mattgaetz close ...|
|1512792456739311624|[saturday, mornin...|          374834303|     FloridaTGA|The FTGA is dedic...|   null|Saturday morning ...|
|1512838819535458307|[smartest, kids, ...|1509279218441474048|  EstuardoPaz12|                null|   null|@kennymxu The sma...|
|1513027591887749125|[high, school, fi...|           60645144|       deba1602|Part of the flock...|   null|@manojmishrasays ...|
|1513929606436380677|[join, grass, roo...|          132673656|homesteadtvmaga|networking and sh...|   null|@GovRonDeSantis J...|
|1513982916363419648|[providence, scho...|          850883642|     ApinCiihuy|                null|   null|Provid
```

```
ence School...|
|1513982956033175553|[academy, benjami...|          850883642|     ApinCiihuy|                null|   null|King's
Academy vs...|
|1513996075434135553|[excited, recruit...|          209220162| iamkathyreaves|Assisting High Sc...|   null|Excite
d to be rec...|
|1514000865757118469|[alright, lemme, ...| 766050497803460608| fugitivemonkey|deal with the ran...|   null|@falle
n_up54 @jbo...|
|1514286349653839879|[university, prof...|1031220149125308418|  whitneyjourno|00's movie stan &...|   null|To the
university...|
|1514353050172936197|[congrats, perfec...|           23967005|HoldernessToday|A community commi...|   null|Congra
ts to Tan '...|
|1514415884114268160|     [math, college]|           14661957|        Mitch_M|Consultant & Spea...|   null|@terri
nakamura @m...|
|1514691503020556292|[high, school, math]|1270451093676404736|     xavieraugs|            OTTERS!! |   null|@verab
eauvoir hig...|
|1514826129198104579|[super, excited, ...|         4252993693|     mega_biggs|final bossbabe | ...|   null|super
excited to ...|
+-------------------+--------------------+-------------------+---------------+--------------------+-------+------
--------------+
only showing top 20 rows
```

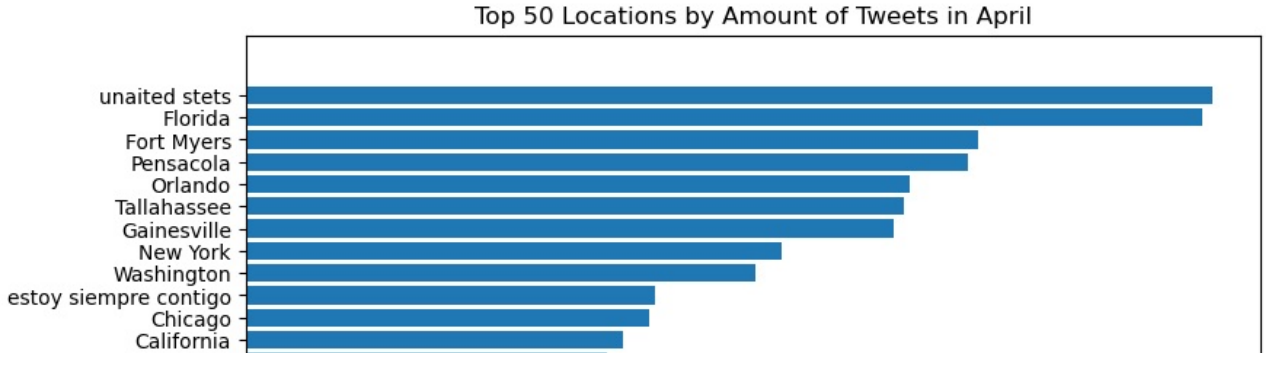
In [74]:
```
tt.count()
```

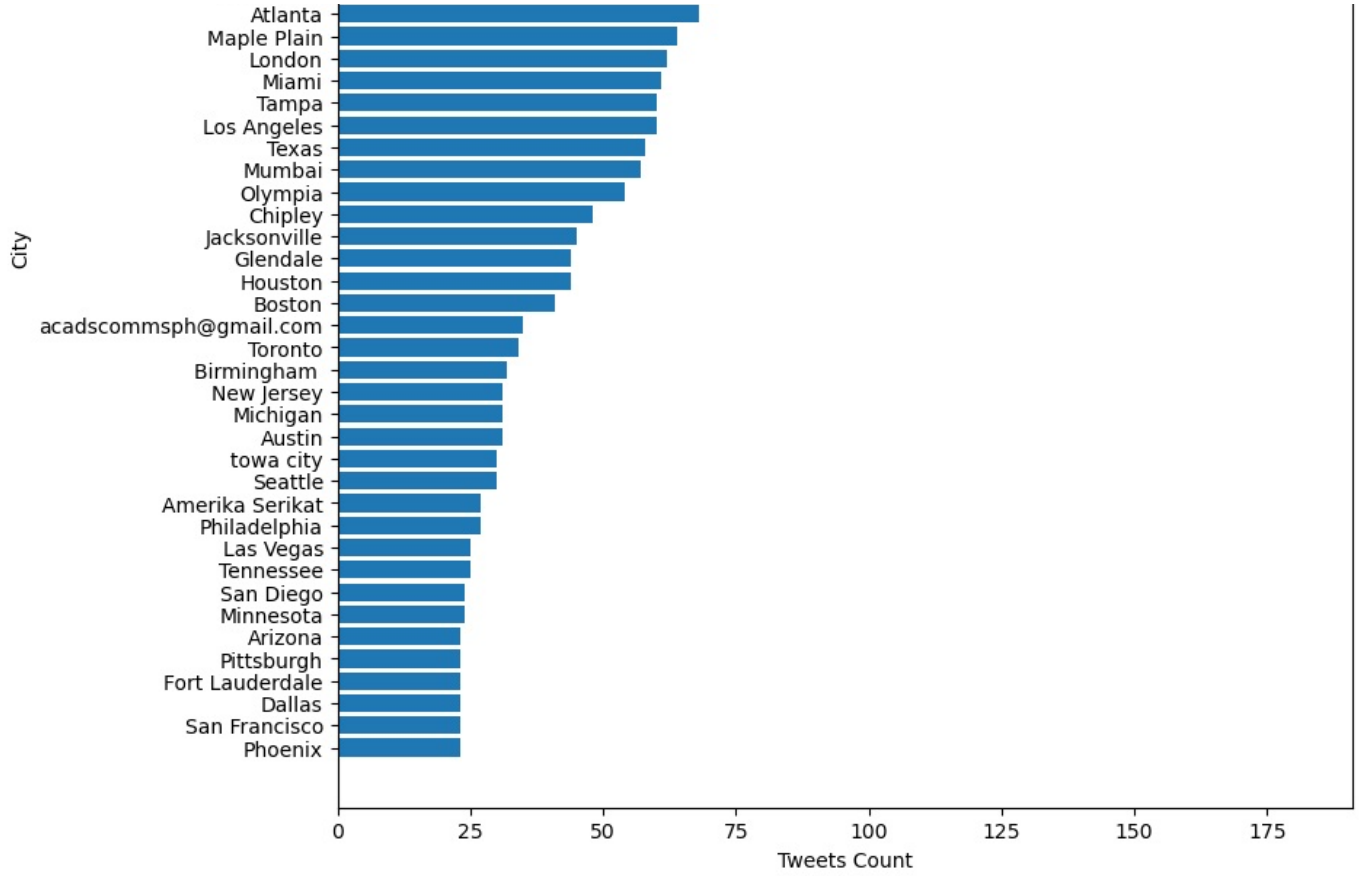
Out[74]: 109509

In [40]:
```
t = tt.withColumn('mnth', substring_index(col('created_at'),' ', 3)).drop('created_at')
t = t.withColumn('mnth', substring_index(col('mnth'),' ',-2))
t = t.withColumn('mnth', substring_index(col('mnth'),' ',1))
t = t.withColumn("mnth",from_unixtime(unix_timestamp(col("mnth"),'MMM'),'MM'))
```

In [33]:
```
april = t.filter(col('mnth')=='04')
may = t.filter(col('mnth')=='05')
june = t.filter(col('mnth')=='06')
july = t.filter(col('mnth')=='07')
aug = t.filter(col('mnth')=='08')
sep = t.filter(col('mnth')=='09')
```

In [34]:
```
loc =april.groupby('user_loc').agg(count('user_loc').alias('loc_ct'))
b = loc.select(substring_index(col('user_loc'),',', 1).alias('loc'),'loc_ct')
b = b.groupby('loc').agg(sum('loc_ct').alias('ct')).sort(col('ct').desc()).limit(50)
location = b.toPandas()
l = ['USA','United States','united states', 'UK','United Kingdom','India','England','South Africa',
     'Nigeria','Canada','she/her', 'Earth']
city = location[~location['loc'].isin(l)]
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(city['loc'],city['ct'])
ax.invert_yaxis()
plt.xlabel('Tweets Count')
plt.ylabel('City')
plt.title('Top 50 Locations by Amount of Tweets in April')
plt.show()
```

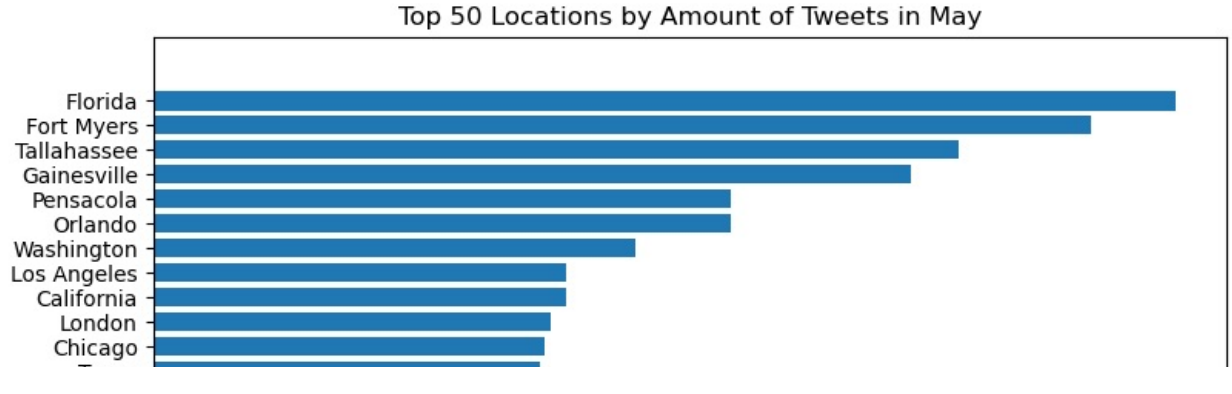


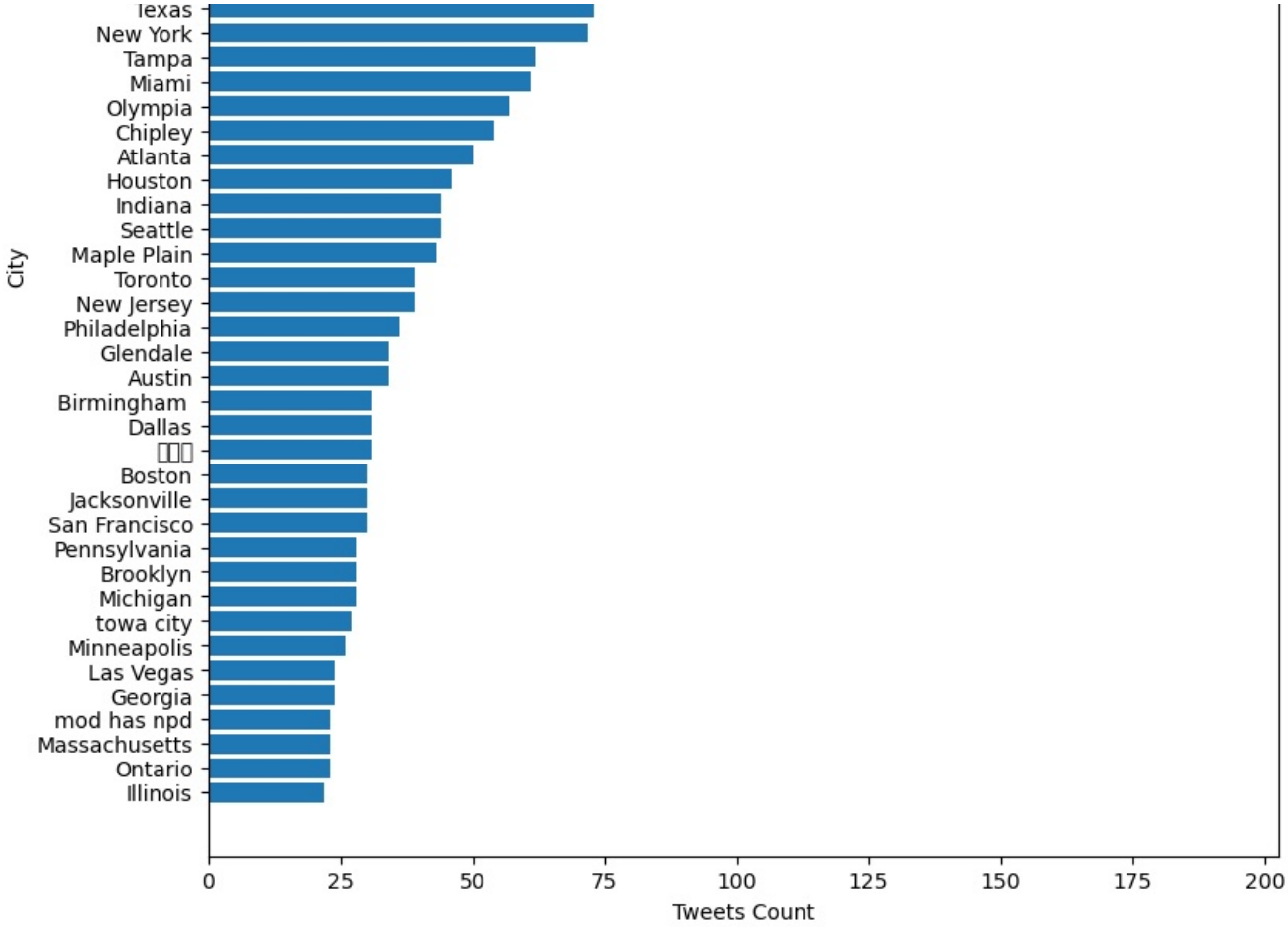



In [35]:

```
loc =may.groupby('user_loc').agg(count('user_loc').alias('loc_ct'))
b = loc.select(substring_index(col('user_loc'),',', 1).alias('loc'),'loc_ct')
b = b.groupby('loc').agg(sum('loc_ct').alias('ct')).sort(col('ct').desc()).limit(50)
location = b.toPandas()
l = ['USA','United States','united states', 'UK','United Kingdom','India','England','South Africa',
     'Nigeria','Canada','she/her', 'Earth']
city = location[~location['loc'].isin(l)]
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(city['loc'],city['ct'])
ax.invert_yaxis()
plt.xlabel('Tweets Count')
plt.ylabel('City')
plt.title('Top 50 Locations by Amount of Tweets in May')
plt.show()
```

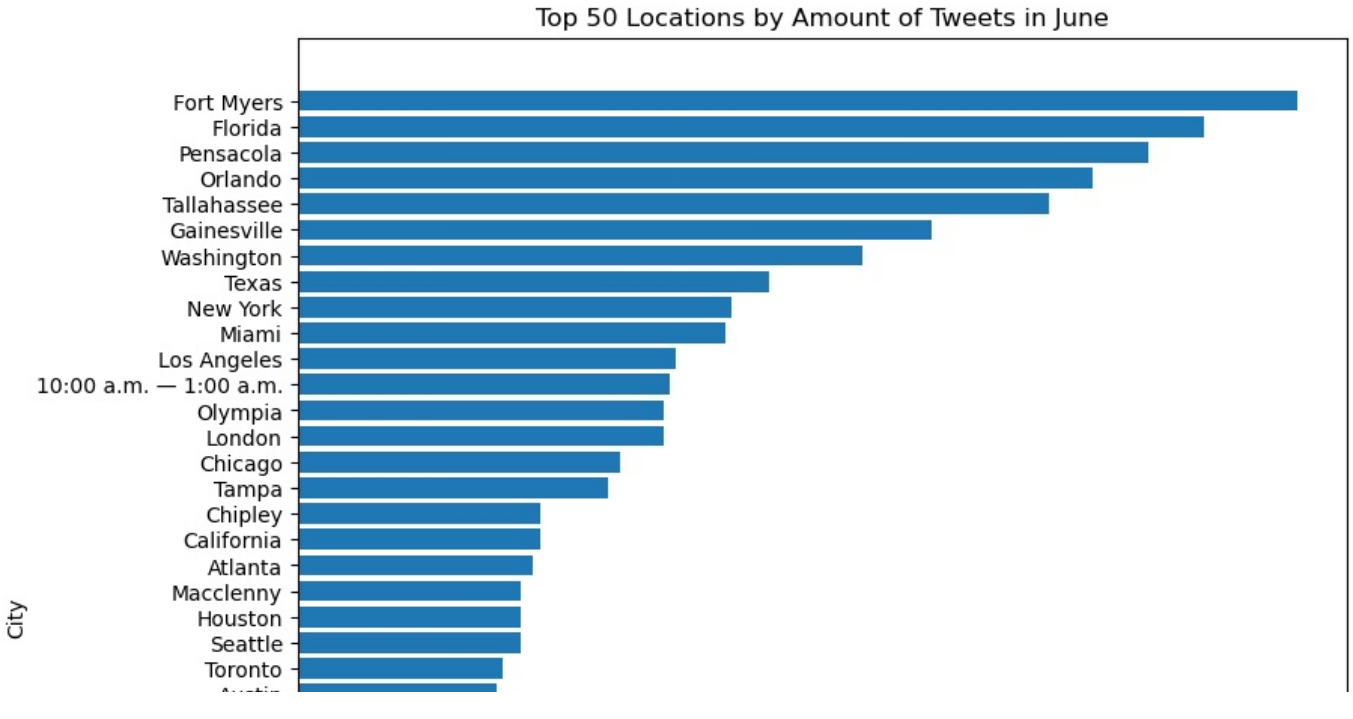
```
/opt/conda/miniconda3/lib/python3.8/site-packages/matplotlib/backends/backend_agg.py:240: RuntimeWarning: Glyph 2
1271 missing from current font.
  font.set_text(s, 0.0, flags=flags)
/opt/conda/miniconda3/lib/python3.8/site-packages/matplotlib/backends/backend_agg.py:240: RuntimeWarning: Glyph 2
8023 missing from current font.
  font.set_text(s, 0.0, flags=flags)
/opt/conda/miniconda3/lib/python3.8/site-packages/matplotlib/backends/backend_agg.py:240: RuntimeWarning: Glyph 3
6947 missing from current font.
  font.set_text(s, 0.0, flags=flags)
/opt/conda/miniconda3/lib/python3.8/site-packages/matplotlib/backends/backend_agg.py:203: RuntimeWarning: Glyph 2
1271 missing from current font.
  font.set_text(s, 0, flags=flags)
/opt/conda/miniconda3/lib/python3.8/site-packages/matplotlib/backends/backend_agg.py:203: RuntimeWarning: Glyph 2
8023 missing from current font.
  font.set_text(s, 0, flags=flags)
/opt/conda/miniconda3/lib/python3.8/site-packages/matplotlib/backends/backend_agg.py:203: RuntimeWarning: Glyph 3
6947 missing from current font.
  font.set_text(s, 0, flags=flags)
```

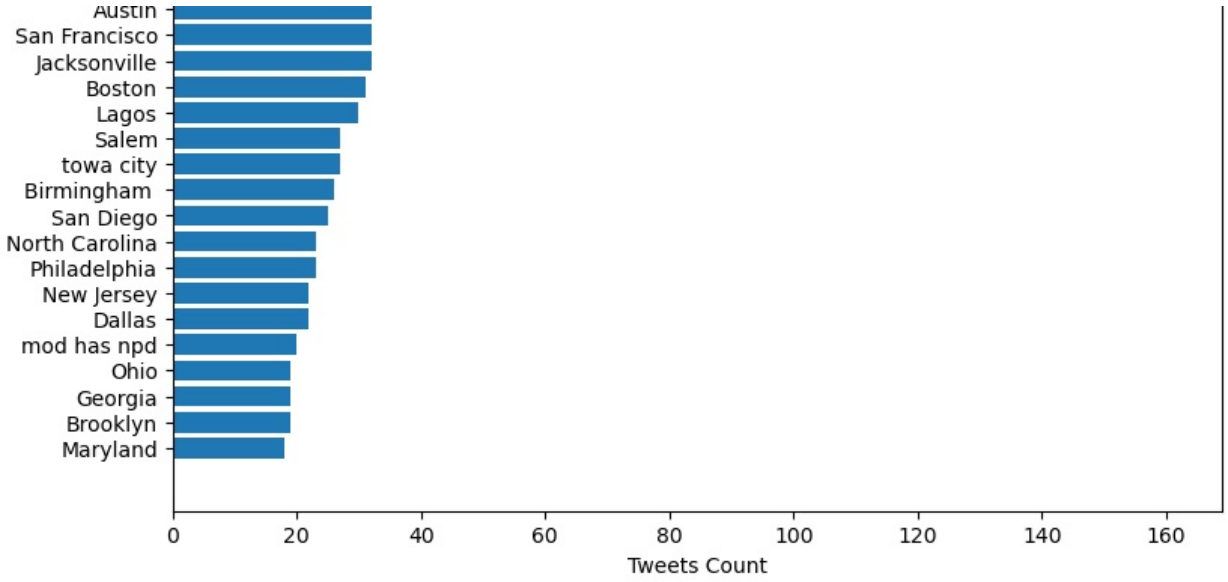





In [36]:

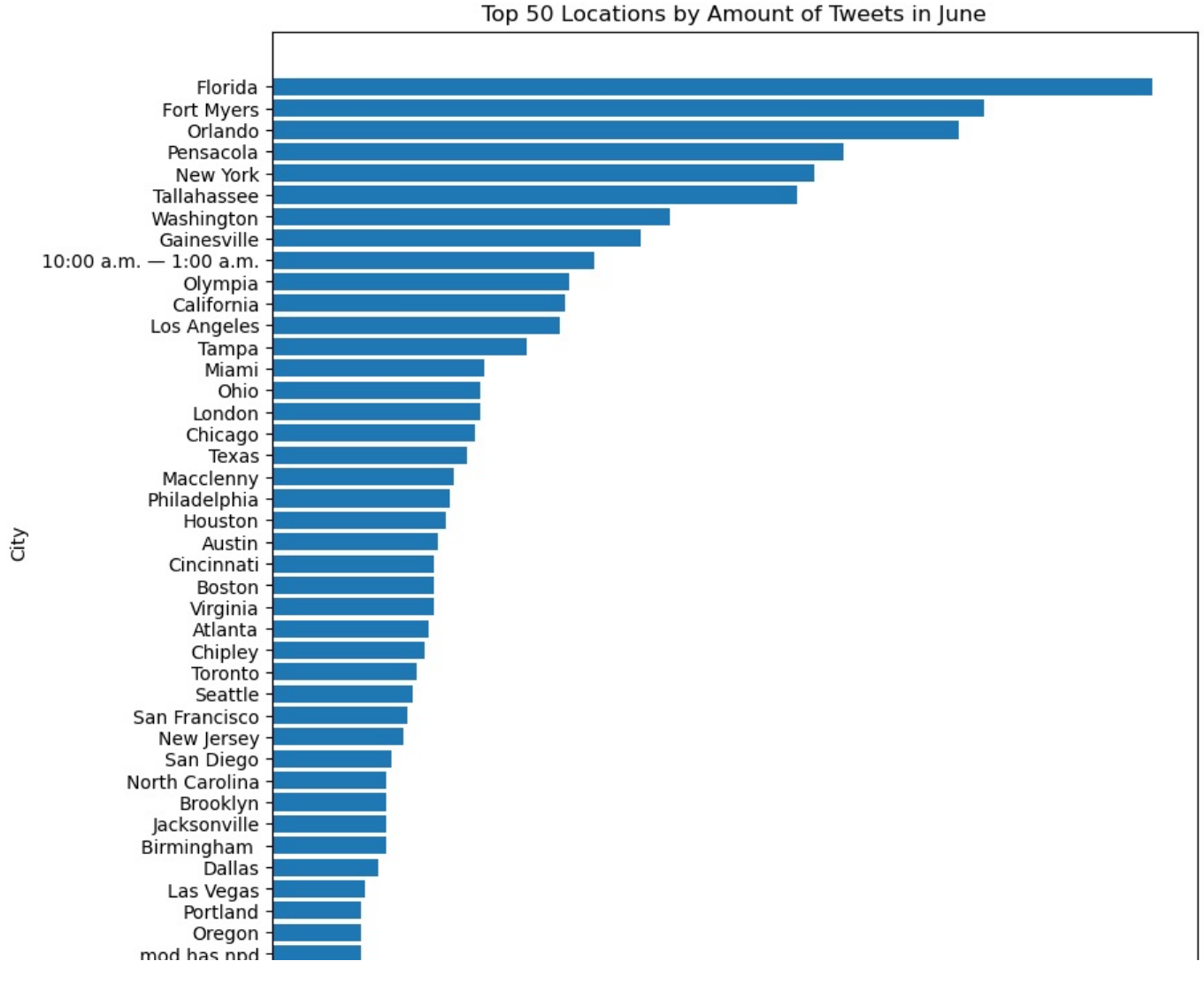
```
loc =june.groupby('user_loc').agg(count('user_loc').alias('loc_ct'))
b = loc.select(substring_index(col('user_loc'),',', 1).alias('loc'),'loc_ct')
b = b.groupby('loc').agg(sum('loc_ct').alias('ct')).sort(col('ct').desc()).limit(50)
location = b.toPandas()
l = ['USA','United States','united states', 'UK','United Kingdom','India','England','South Africa',
     'Nigeria','Canada','she/her', 'Earth']
city = location[~location['loc'].isin(l)]
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(city['loc'],city['ct'])
ax.invert_yaxis()
plt.xlabel('Tweets Count')
plt.ylabel('City')
plt.title('Top 50 Locations by Amount of Tweets in June')
plt.show()
```

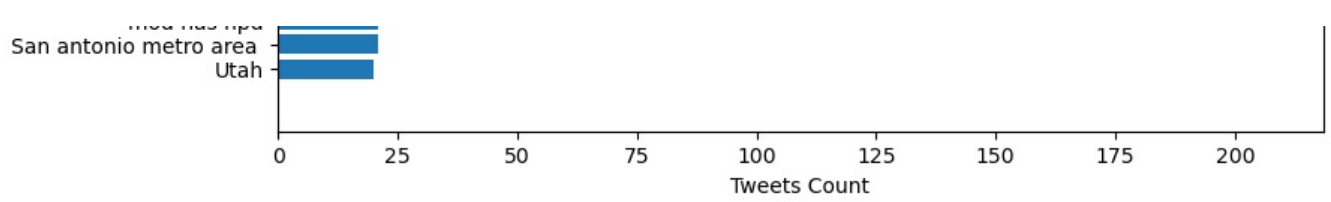





In [48]:

```
loc =july.groupby('user_loc').agg(count('user_loc').alias('loc_ct'))
b = loc.select(substring_index(col('user_loc'),',', 1).alias('loc'),'loc_ct')
b = b.groupby('loc').agg(sum('loc_ct').alias('ct')).sort(col('ct').desc()).limit(50)
location = b.toPandas()
l = ['USA','United States','united states', 'UK','United Kingdom','India','England','South Africa',
     'Nigeria','Canada','she/her', 'Earth']
city = location[~location['loc'].isin(l)]
fig, ax = plt.subplots(figsize =(9, 10))
ax.barh(city['loc'],city['ct'])
ax.invert_yaxis()
plt.xlabel('Tweets Count')
plt.ylabel('City')
plt.title('Top 50 Locations by Amount of Tweets in July')
plt.show()
```






## Timelines

In [137…

```
a = text.select('created_at')
a = a.select(substring_index(col('created_at'),' ', 3).alias('date'))
a = a.select(substring_index(col('date'),' ',-2).alias('m/d'))
a = a.select(substring_index(col('m/d'),' ',1).alias('Month'))
a = a.withColumn("mnth",from_unixtime(unix_timestamp(col("Month"),'MMM'),'MM')).select('mnth')
a = a.groupby('mnth').agg(count('mnth'))
a.show()
```

```
+----+-----------+
|mnth|count(mnth)|
+----+-----------+
|  07|    3256544|
|  11|     928453|
|  09|    4677706|
|  05|    4245053|
|  08|    4221132|
|  06|    3675218|
|  10|    4678795|
|  04|    3445961|
+----+-----------+
```

In [141…

```
a = text.select('created_at')
a = a.select(substring_index(col('created_at'),' ', 3).alias('date'))
a = a.select(substring_index(col('date'),' ',-2).alias('m/d'))
a = a.select(substring_index(col('m/d'),' ',1).alias('Month'))
a = a.withColumn("mnth",from_unixtime(unix_timestamp(col("Month"),'MMM'),'MM')).select('mnth')
a = a.groupby('mnth').agg(count('mnth').alias('count'))
```

In [144…

```
a = a.withColumn("month",to_timestamp("mnth", 'MM'))
a = a.withColumn("Mon", date_format(col("month"), "M"))
a = a.select('Mon', 'count')
time = a.toPandas()
```

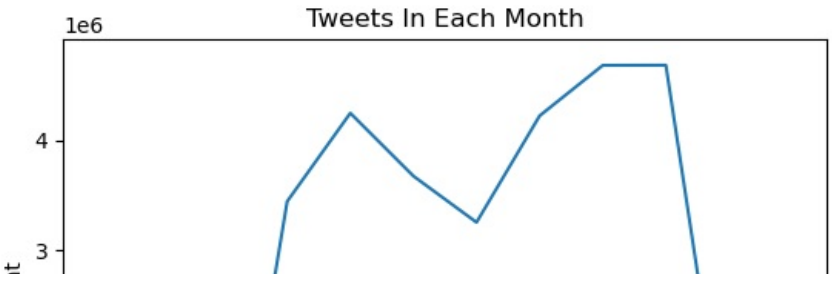
In [157…

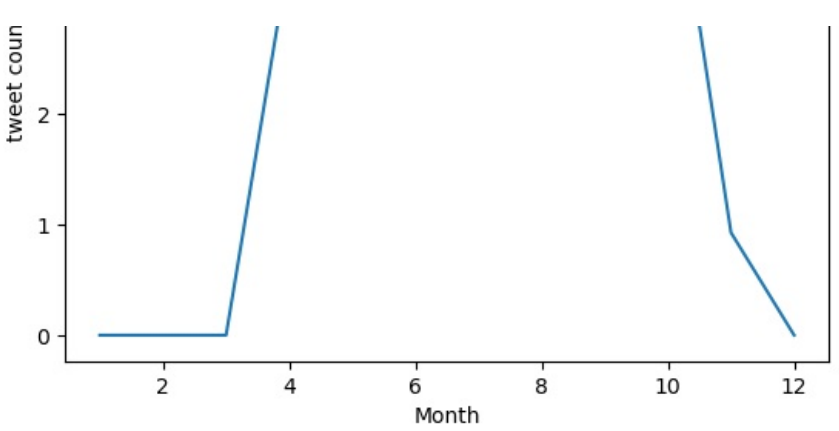
```
missing = pd.DataFrame({'Mon':['1','2','3','12'], 'count':[0,0,0,0]})
time = time.append(missing)
time['Mon'] = time['Mon'].astype(int)
time = time.sort_values(by = ['Mon'])
```

In [163…

```
plt.plot(time['Mon'], time['count'])
plt.xlabel('Month')
plt.ylabel('tweet count')
plt.title('Tweets In Each Month')
```

Out[163… Text(0.5, 1.0, 'Tweets In Each Month')





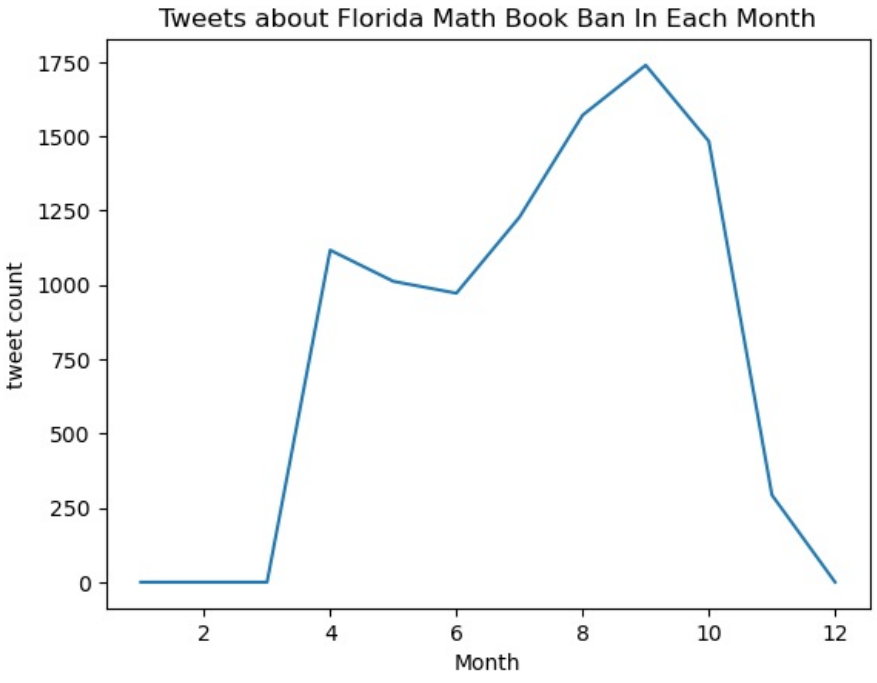

In [ ]:
```
# Does not have any records in January, February, March, December, and only few records in November
# Peaks are mostly happen in Autumn and Spring
```

In [44]:
```
### topic timeline
a = t.groupby('mnth').agg(count('mnth').alias('count'))
a = a.withColumn("month",to_timestamp("mnth", 'MM'))
a = a.withColumn("Mon", date_format(col("month"), "M"))
a = a.select('Mon', 'count')
time = a.toPandas()
```

In [45]:
```
missing = pd.DataFrame({'Mon':['1','2','3','12'], 'count':[0,0,0,0]})
time = time.append(missing)
time['Mon'] = time['Mon'].astype(int)
time = time.sort_values(by = ['Mon'])
```

In [47]:
```
plt.plot(time['Mon'], time['count'])
plt.xlabel('Month')
plt.ylabel('tweet count')
plt.title('Tweets about Florida Math Book Ban In Each Month')
```

Out[47]: Text(0.5, 1.0, 'Tweets about Florida Math Book Ban In Each Month')




## Uniqueness of Tweets

In [4]:
```
import pandas as pd
```

```
import pandas as pd
import matplotlib.pyplot as plt
from pyspark.sql.functions import *
text = spark.read.parquet('gs://'+'msca-bdp-students-bucket/shared_data/chenfeng/project/key_eda/')
original = text.filter(col('retweet').isNull())
original = original.drop('tokens')
```

In [5]:

```
from pyspark.sql import functions as F
from pyspark.sql import types as t
from pyspark.sql.types import ArrayType, IntegerType, BooleanType

eng_ord=F.udf(lambda x: [ord(a) for a in x],t.ArrayType(IntegerType()))
def eng_filter(x):
    for index in range(len(x)):
        if x[index] > 128:
            return False
    return True

filter_udf = F.udf(eng_filter, BooleanType())
eng_ori = original.filter(filter_udf(eng_ord('full_text')) == True)
```

In [6]:

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

d = eng_ori.rdd.map(lambda x : x['full_text']).filter(lambda x: x is not None)
StopWords = stopwords.words("english")
# remove stop words
tokens = d\
    .map( lambda document: document.strip().lower())\
    .map( lambda document: re.sub("@[A-Za-z0-9_]+","", document))\
    .map( lambda document: re.sub(r'[^\w\s]', '', document))\
    .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])
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [10]:

```
t = tokens.zip(eng_ori.select('id').rdd.flatMap(lambda x:x))
df_tokens = t.toDF(['tokens', 'id'])
eng = df_tokens.filter(size("tokens")>=1)
```

In [11]:

```
eng.count()
```

Out[11]: 7103950

In [13]:

```
original.count()
```

Out[13]: 11841802

In [76]:

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

In [17]:

```
eng = spark.read.parquet('gs://'+'msca-bdp-students-bucket/shared_data/chenfeng/project/filtered/')
```

In [48]:
```
sample = eng.limit(10000)
```

In [17]:
```
sample.show()
```

```
[Stage 11:=============================================================>(66 + 1) / 67]
```

```
+--------------------+------------------+
|              tokens|                id|
+--------------------+------------------+
|[size, nike, forc...|1511200232385613828|
|[love, living, co...|1511204236956254213|
|[watch, live, web...|1511208753114361856|
|[weather, summary...|1511210266352132098|
|[sound, thinking,...|1511213904931741699|
|[2022, high, scho...|1511220002963214341|
|[college, basketb...|1511220059183599618|
| [college, students]|1511220675352145920|
|[asad, meraj, pri...|1511224190028095490|
|[biosciences, sum...|1511229778757664771|
|[minority, univer...|1511230052553588736|
|[grants, wisconsi...|1511232354626543620|
|[bear, keychain, ...|1511235831696355330|
|[looking, custom,...|1511238105650667522|
|[delhi, college, ...|1511239927132200964|
|[remembered, frie...|1511240014654496771|
|[medical, college...|1511244257927249920|
|[january, kwania,...|1511245001312948226|
|[missing, darwin,...|1511249972737499136|
|[basic, education...|1511253275445641219|
+--------------------+------------------+
only showing top 20 rows
```

In [49]:
```
import re
from pyspark.ml.feature import MinHashLSH
from pyspark.ml.feature import CountVectorizer,  IDF, CountVectorizerModel, Tokenizer, RegexTokenizer, StopWordsR
vectorize = CountVectorizer(inputCol="tokens", outputCol="features", minDF=1.0)
text_vectorize = vectorize.fit(sample).transform(sample)
mh = MinHashLSH(inputCol="features", outputCol="hashes", numHashTables=5)
text_model = mh.fit(text_vectorize)
text_hashed = mh.fit(text_vectorize).transform(text_vectorize)
```

In [50]:
```
jaccard_distance = 0.5

df_dups_text = text_model.approxSimilarityJoin(text_hashed, text_hashed, jaccard_distance).\
               filter("datasetA.id < datasetB.id").select(col("distCol"),\
            col("datasetA.id").alias("id_A"),
           col("datasetB.id").alias("id_B")
          )
df_dups_text.show()
```

```
[Stage 133:>                                                        (0 + 1) / 1]
```

```
+------------------+------------------+------------------+
|           distCol|              id_A|              id_B|
+------------------+------------------+------------------+
|0.3333333333333337|1511198670347292672|1514745869186342920|
|               0.0|1511200011505532928|1511604982990372865|
|               0.0|1511200109303803904|1511202334759555072|
|               0.0|1511200619562147840|1511305941345456129|
|               0.0|1511200619562147840|1511209914919981062|
|              0.25|1511201546264055811|1514011148668047363|
|              0.25|1511201546264055811|1513843596637655040|
|               0.4|1511201546264055811|1511369299931414537|
|               0.0|1511203962745196544|1511317321670959105|
|               0.0|1511203962745196544|1511315593504378882|
|               0.0|1511203962745196544|1511302925741613057|
|               0.0|1511203962745196544|1511293945774489601|
```

```
|                0.0|1511203962745196544|1511259973954809864|
|                0.0|1511203962745196544|1511246097183739907|
|                0.0|1511203962745196544|1511242738599084038|
|                0.0|1511203962745196544|1511219386815635459|
|                0.0|1511203962745196544|1511217033823887367|
|                0.0|1511204174221877248|1511317321670959105|
|                0.0|1511204174221877248|1511315593504378882|
|                0.0|1511204174221877248|1511302925741613057|
+-------------------+-------------------+-------------------+
only showing top 20 rows
```

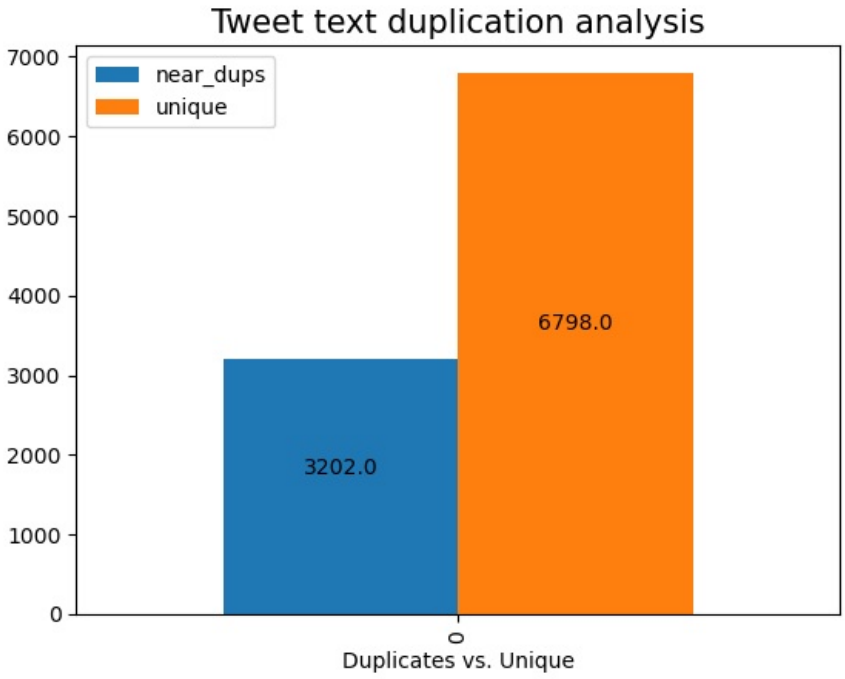
In [51]:

```
records = text_hashed.count()
dups = df_dups_text.select('id_A').distinct().count()
dups_50 = dups
uniques = records - dups

print ('Total records: ', records)
print ('Duplicate text based on {', jaccard_distance, '} jaccard distance: ', dups)
print ('Unique text based on {', jaccard_distance, '} jaccard distance: ', jaccard_distance, ': ', uniques)
```

```
[Stage 141:>                                                        (0 + 1) / 1]
Total records:  10000
Duplicate text based on { 0.5 } jaccard distance:  3202
Unique text based on { 0.5 } jaccard distance:  0.5 :  6798
```

In [52]:

```
dups_df = pd.DataFrame.from_dict({'near_dups': [dups], 'unique': [uniques]})

ax=dups_df.plot(kind = 'bar',y=['near_dups', 'unique'], fontsize=10, color=['C0', 'C1'], align='center', width=0.
                xlabel="Duplicates vs. Unique")
ax.set_title('Tweet text duplication analysis', fontsize=15)
for p in ax.patches:
       ax.annotate(format(p.get_height(), '.1f'),
                   (p.get_x() + p.get_width() / 2., p.get_height()/2),
                   ha = 'center', va = 'center',
                   xytext = (0, 9),
                   textcoords = 'offset points')
```




In [18]:

```
df_dups_text.cache()
```

Out[18]:

```
DataFrame[distCol: double, id_A: bigint, id_B: bigint]
```

In [19]:
```
ori_text = original.select('id','full_text')
df = df_dups_text.join(ori_text, df_dups_text.id_A == ori_text.id).select(df_dups_text.id_A, 'full_text', 'id_B')
df = df.withColumnRenamed('full_text', 'full_text_A')
df = df.join(ori_text, df_dups_text.id_B == ori_text.id).drop(ori_text.id)
df = df.withColumnRenamed('full_text', 'full_text_B')
```

In [20]:
```
df_pd = df.toPandas()
df_pd
```

Out[20]:

| | id_A | full_text_A | id_B | full_text_B |
|---|---|---|---|---|
| **0** | 1511335225225363460 | @hazelAhmady @alexplitsas @shawnjvandiver Stud... | 1511349143251427332 | @SalemNowrozie Students of the American Univer... |
| **1** | 1511335225225363460 | @hazelAhmady @alexplitsas @shawnjvandiver Stud... | 1511397631217516547 | @TheCryptoLark Students of the American Univer... |
| **2** | 1511335225225363460 | @hazelAhmady @alexplitsas @shawnjvandiver Stud... | 1511401925148913664 | @RohaNadym Students of the American University... |
| **3** | 1511335225225363460 | @hazelAhmady @alexplitsas @shawnjvandiver Stud... | 1511407790811758597 | @JDCocchiarella Students of the American Unive... |
| **4** | 1511459323666628609 | Gilmour Academy vs Hawken \| Ohio High School B... | 1511465633820856321 | Hubbard vs Jefferson Area \| Ohio High School B... |
| **...** | ... | ... | ... | ... |
| **332** | 1511625079586975744 | @GabbTineOFC Do your laundry\n\nGABBTINE Unive... | 1513460851646341121 | @SCynic1 @MartinDaubney @nusuk What university... |
| **333** | 1511625463780716549 | @GabbTineOFC bake a cake\n\nGABBTINE University | 1513460851646341121 | @SCynic1 @MartinDaubney @nusuk What university... |
| **334** | 1511685838161260545 | @RadharamnDas @myogiadityanath @PMOIndia This ... | 1513460851646341121 | @SCynic1 @MartinDaubney @nusuk What university... |
| **335** | 1512158275315265540 | Agile Leadership\n#Agile #Leadership\nStanford... | 1513460851646341121 | @SCynic1 @MartinDaubney @nusuk What university... |
| **336** | 1512518146112069644 | Did you go to the University of Arizona? | 1513460851646341121 | @SCynic1 @MartinDaubney @nusuk What university... |

337 rows × 4 columns

In [21]:
```
pd.set_option('display.max_rows', 30)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('max_colwidth', -1)
df_pd[['full_text_A','full_text_B']]
```

```
/tmp/ipykernel_29030/1424795411.py:3: FutureWarning: Passing a negative integer is deprecated in version 1.0 and
will not be supported in future version. Instead, use None to not limit the column width.
  pd.set_option('max_colwidth', -1)
```

Out[21]:

| | full_text_A | full_text_B |
|---|---|---|
| **0** | @hazelAhmady @alexplitsas @shawnjvandiver Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents | @SalemNowrozie Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents |
| **1** | @hazelAhmady @alexplitsas @shawnjvandiver Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents | @TheCryptoLark Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents |
| **2** | @hazelAhmady @alexplitsas @shawnjvandiver Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents | @RohaNadym Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents |
| **3** | @hazelAhmady @alexplitsas @shawnjvandiver Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents | @JDCocchiarella Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents |
| **4** | Gilmour Academy vs Hawken \| Ohio High School Baseball 2022 Live Stream \| Click Here Watch Live HD ==> https://t.co/FFzjlbV6Cr | Hubbard vs Jefferson Area \| Ohio High School Baseball 2022 Live Stream \| Click Here Watch Live HD ==> https://t.co/g2bJT4OKvM |
| **...** | ... | ... |
| **332** | @GabbTineOFC Do your laundry\n\nGABBTINE University | @SCynic1 @MartinDaubney @nusuk What university is this at? |
| **333** | @GabbTineOFC bake a cake\n\nGABBTINE University | @SCynic1 @MartinDaubney @nusuk What university is this at? |
| | @RadharamnDas @myogiadityanath @PMOIndia This university itself should | |

| | | |
|---|---|---|
| 334 | cease to exist now. | @SCynic1 @MartinDaubney @nusuk What university is this at? |
| 335 | Agile Leadership\n#Agile #Leadership\nStanford University & University of Pennsylvania\n-> https://t.co/tqfHjxQ0sG | @SCynic1 @MartinDaubney @nusuk What university is this at? |
| 336 | Did you go to the University of Arizona? | @SCynic1 @MartinDaubney @nusuk What university is this at? |

337 rows × 2 columns

In [ ]:
```
#### The jaccard similarity distance for 0.5 seems a little too high since the actual text looks different
```

In [39]:
```
sample = eng.limit(10000)
```

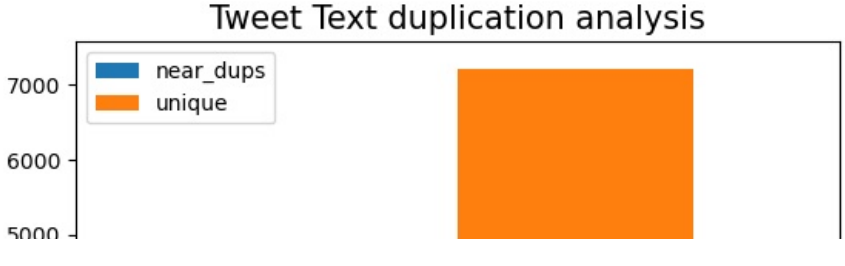
In [42]:
```
vectorize = CountVectorizer(inputCol="tokens", outputCol="features", minDF=1.0)
text_vectorize = vectorize.fit(sample).transform(sample)
mh = MinHashLSH(inputCol="features", outputCol="hashes", numHashTables=5)
text_model = mh.fit(text_vectorize)
text_hashed = mh.fit(text_vectorize).transform(text_vectorize)
```

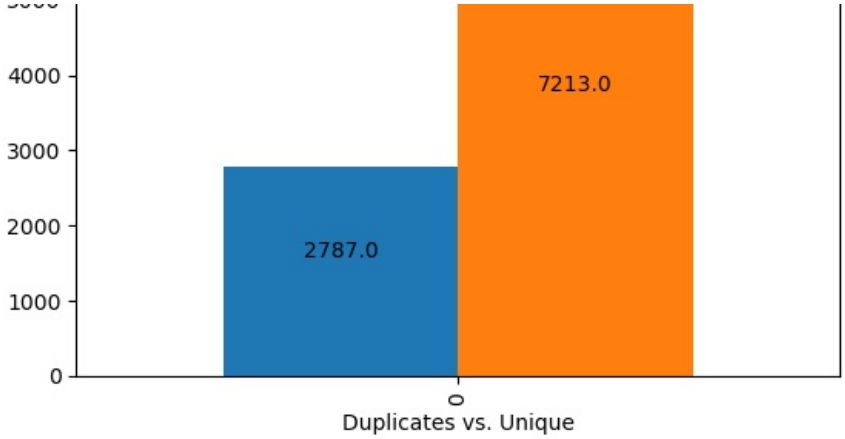
In [43]:
```
jaccard_distance = 0.20

df_dups_text = text_model.approxSimilarityJoin(text_hashed, text_hashed, jaccard_distance).\
                filter("datasetA.id < datasetB.id").select(col("distCol"),\
             col("datasetA.id").alias("id_A"),
            col("datasetB.id").alias("id_B")
           )
```

In [44]:
```
records = text_hashed.count()
dups = df_dups_text.select('id_A').distinct().count()
dups_50 = dups
uniques = records - dups

print ('Total records: ', records)
print ('Duplicate text based on {', jaccard_distance, '} jaccard distance: ', dups)
print ('Unique text based on {', jaccard_distance, '} jaccard distance: ', jaccard_distance, ': ', uniques)
```

```
22/12/05 03:38:09 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'.
[Stage 121:>                                                        (0 + 1) / 1]
Total records:  10000
Duplicate text based on { 0.2 } jaccard distance:  2787
Unique text based on { 0.2 } jaccard distance:  0.2 :  7213
```

In [46]:
```
dups_df = pd.DataFrame.from_dict({'near_dups': [dups], 'unique': [uniques]})

ax=dups_df.plot(kind = 'bar',y=['near_dups', 'unique'], fontsize=10, color=['C0', 'C1'], align='center', width=0.
                xlabel="Duplicates vs. Unique")
ax.set_title('Tweet Text duplication analysis', fontsize=15)
for p in ax.patches:
       ax.annotate(format(p.get_height(), '.1f'),
                   (p.get_x() + p.get_width() / 2., p.get_height()/2),
                   ha = 'center', va = 'center',
                   xytext = (0, 9),
                   textcoords = 'offset points')
```






In [30]:
```
ori_text = original.select('id','full_text')
df = df_dups_text.join(ori_text, df_dups_text.id_A == ori_text.id).select(df_dups_text.id_A, 'full_text', 'id_B')
df = df.withColumnRenamed('full_text', 'full_text_A')
df = df.join(ori_text, df_dups_text.id_B == ori_text.id).drop(ori_text.id)
df = df.withColumnRenamed('full_text', 'full_text_B')
df_pd = df.toPandas()
```

In [31]:
```
pd.set_option('display.max_rows', 30)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('max_colwidth', -1)
df_pd[['full_text_A','full_text_B']]
```

```
/tmp/ipykernel_29030/1424795411.py:3: FutureWarning: Passing a negative integer is deprecated in version 1.0 and
will not be supported in future version. Instead, use None to not limit the column width.
  pd.set_option('max_colwidth', -1)
```

Out[31]:

| | full_text_A | full_text_B |
|---|---|---|
| 0 | @WHCOS Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents | @SkyNews Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents |
| 1 | @WHCOS Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents | @TristanSnell Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents |
| 2 | @SkyNews Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents | @TristanSnell Students of the American University of Afghanistan are in grave danger and their future is uncertain. Please evacuate them! So they continue their education without any barrier and fear! Thank you \n\n#EvacuateAUAFStudents |
| 3 | @SenSanders Reverse the 1981 decision that cut benefits for survival SSA benefits at 12th grade College cost have skyrocketed and grants have NOT kept up They go deep into lifelong student loans and #CancelStudentDebt over 200,000 lost a parent to Covid | @YoungInvincible @POTUS Reverse the 1981 decision that cut benefits for survival SSA benefits at 12th grade College cost have skyrocketed and grants have NOT kept up They go deep into lifelong student loans and #CancelStudentDebt over 200,000 lost a parent to Covid |
| 4 | @SenSanders Reverse the 1981 decision that cut benefits for survival SSA benefits at 12th grade College cost have skyrocketed and grants have NOT kept up They go deep into lifelong student loans and #CancelStudentDebt over 200,000 lost a parent to Covid | @RBReich Robert Reich please advocate this 1981 decision that cut benefits for survival SSA benefits at 12th grade College cost have skyrocketed and grants have NOT kept up They go deep into lifelong student loans and #CancelStudentDebt over 200,000 lost a parent to Covid |
| ... | ... | ... |
| 113 | @RobertToy1 @daniellamyoung You were in college in 1981?! | Pandemonium ensues as pine tar overturns college HR https://t.co/vaEx9LO9QU via @mlb |
| 114 | Ahhhh college *wistfully* | Pandemonium ensues as pine tar overturns college HR https://t.co/vaEx9LO9QU via @mlb |
| 115 | College in 2 sentences | Pandemonium ensues as pine tar overturns college HR https://t.co/vaEx9LO9QU via @mlb |
| 116 | Just Enrolledn in Rat college | Pandemonium ensues as pine tar overturns college HR https://t.co/vaEx9LO9QU via @mlb |
| 117 | college was such a mistake | Pandemonium ensues as pine tar overturns college HR https://t.co/vaEx9LO9QU via @mlb |

118 rows × 2 columns

In [33]:
```
ori_text = eng.select('id','tokens')
```

```
df = df_dups_text.join(ori_text, df_dups_text.id_A == ori_text.id).select(df_dups_text.id_A, 'tokens', 'id_B')
df = df.withColumnRenamed('tokens', 'tokens_A')
df = df.join(ori_text, df_dups_text.id_B == ori_text.id).drop(ori_text.id)
df = df.withColumnRenamed('tokens', 'tokens_B')
df_pd = df.toPandas()
```

```
22/12/05 03:23:21 WARN org.apache.spark.network.server.TransportChannelHandler: Exception in connection from /10.
128.0.249:40382
java.io.IOException: Connection timed out
        at sun.nio.ch.FileDispatcherImpl.read0(Native Method)
        at sun.nio.ch.SocketDispatcher.read(SocketDispatcher.java:39)
        at sun.nio.ch.IOUtil.readIntoNativeBuffer(IOUtil.java:223)
        at sun.nio.ch.IOUtil.read(IOUtil.java:192)
        at sun.nio.ch.SocketChannelImpl.read(SocketChannelImpl.java:379)
        at io.netty.buffer.PooledByteBuf.setBytes(PooledByteBuf.java:253)
        at io.netty.buffer.AbstractByteBuf.writeBytes(AbstractByteBuf.java:1133)
        at io.netty.channel.socket.nio.NioSocketChannel.doReadBytes(NioSocketChannel.java:350)
        at io.netty.channel.nio.AbstractNioByteChannel$NioByteUnsafe.read(AbstractNioByteChannel.java:148)
        at io.netty.channel.nio.NioEventLoop.processSelectedKey(NioEventLoop.java:714)
        at io.netty.channel.nio.NioEventLoop.processSelectedKeysOptimized(NioEventLoop.java:650)
        at io.netty.channel.nio.NioEventLoop.processSelectedKeys(NioEventLoop.java:576)
        at io.netty.channel.nio.NioEventLoop.run(NioEventLoop.java:493)
        at io.netty.util.concurrent.SingleThreadEventExecutor$4.run(SingleThreadEventExecutor.java:989)
        at io.netty.util.internal.ThreadExecutorMap$2.run(ThreadExecutorMap.java:74)
        at io.netty.util.concurrent.FastThreadLocalRunnable.run(FastThreadLocalRunnable.java:30)
        at java.lang.Thread.run(Thread.java:750)
22/12/05 03:23:21 WARN org.apache.spark.network.server.TransportChannelHandler: Exception in connection from /10.
128.0.249:40374
java.io.IOException: Connection timed out
        at sun.nio.ch.FileDispatcherImpl.read0(Native Method)
        at sun.nio.ch.SocketDispatcher.read(SocketDispatcher.java:39)
        at sun.nio.ch.IOUtil.readIntoNativeBuffer(IOUtil.java:223)
        at sun.nio.ch.IOUtil.read(IOUtil.java:192)
        at sun.nio.ch.SocketChannelImpl.read(SocketChannelImpl.java:379)
        at io.netty.buffer.PooledByteBuf.setBytes(PooledByteBuf.java:253)
        at io.netty.buffer.AbstractByteBuf.writeBytes(AbstractByteBuf.java:1133)
        at io.netty.channel.socket.nio.NioSocketChannel.doReadBytes(NioSocketChannel.java:350)
        at io.netty.channel.nio.AbstractNioByteChannel$NioByteUnsafe.read(AbstractNioByteChannel.java:148)
        at io.netty.channel.nio.NioEventLoop.processSelectedKey(NioEventLoop.java:714)
        at io.netty.channel.nio.NioEventLoop.processSelectedKeysOptimized(NioEventLoop.java:650)
        at io.netty.channel.nio.NioEventLoop.processSelectedKeys(NioEventLoop.java:576)
        at io.netty.channel.nio.NioEventLoop.run(NioEventLoop.java:493)
        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)
```

In [38]:

```
pd.set_option('display.max_rows', None)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('max_colwidth', -1)
df_pd[['tokens_A','tokens_B']]
```

```
/tmp/ipykernel_29030/2173607983.py:3: FutureWarning: Passing a negative integer is deprecated in version 1.0 and
will not be supported in future version. Instead, use None to not limit the column width.
  pd.set_option('max_colwidth', -1)
```

Out[38]:

| | tokens_A | tokens_B |
|---|---|---|
| 0 | [students, american, university, afghanistan, grave, danger, future, please, evacuate, continue, education, without, barrier, thank] | [students, american, university, afghanistan, grave, danger, future, please, evacuate, continue, education, without, barrier, thank] |
| 1 | [grand, magus, bellgrove, university] | [university, santo, tomas] |
| 2 | [grand, magus, bellgrove, university] | [university] |
| 3 | [university, santo, tomas] | [university] |
| 4 | [grand, magus, bellgrove, university] | [university] |
| 5 | [university, santo, tomas] | [university] |
| 6 | [university] | [university] |
| 7 | [grand, magus, bellgrove, university] | [bake, university] |
| 8 | [university, santo, tomas] | [bake, university] |
| 9 | [university] | [bake, university] |
| 10 | [university] | [bake, university] |
| 11 | [grand, magus, bellgrove, university] | [university, cease, exist] |
| 12 | [university, santo, tomas] | [university, cease, exist] |

| | | |
|---|---|---|
| **13** | [university] | [university, cease, exist] |
| **14** | [university] | [university, cease, exist] |
| **15** | [bake, university] | [university, cease, exist] |
| **16** | [benjamin, russell, wetumpka, high, school, softball, live, live, wetumpka, varsity, softball, team, home, conference, game, benjamin, russell, today] | [swansea, high, school, softball, live, live, swansea, varsity, softball, team, home, conference, game, today] |
| **17** | [swansea, high, school, softball, live, live, swansea, varsity, softball, team, home, conference, game, today] | [creswell, junction, city, high, school, softball, live, live, junction, city, varsity, softball, team, home, game, creswell, today] |
| **18** | [magazine, identified, college, prospect, help, secure, scholarship, complete, recruiting, profile, follow, coach, scott] | [magazine, identified, college, prospect, help, secure, scholarship, complete, recruiting, profile, follow, coach, scott] |
| **19** | [magazine, identified, college, prospect, help, secure, scholarship, complete, recruiting, profile, follow, coach, scott] | [magazine, identified, college, prospect, help, secure, scholarship, complete, recruiting, profile, follow, coach, scott] |
| **20** | [magazine, identified, college, prospect, help, secure, scholarship, complete, recruiting, profile, follow, coach, scott] | [magazine, identified, college, prospect, help, secure, scholarship, complete, recruiting, profile, follow, coach, scott] |
| **21** | [high, school, janet] | [memba, nigga, high, school] |
| **22** | [high, school, janet] | [high, school, flight, school] |
| **23** | [memba, nigga, high, school] | [high, school, flight, school] |
| **24** | [grand, magus, bellgrove, university] | [agile, university, university] |
| **25** | [university, santo, tomas] | [agile, university, university] |
| **26** | [university] | [agile, university, university] |
| **27** | [university] | [agile, university, university] |
| **28** | [bake, university] | [agile, university, university] |
| **29** | [university, cease, exist] | [agile, university, university] |
| **30** | [memorial, fpca, georgia, high, school, baseball, 2022, live, stream, click, watch, live] | [warner, robins, coffee, georgia, high, school, baseball, 2022, live, stream, click, watch, live] |
| **31** | [madison, county, cedar, shoals, georgia, high, school, baseball, 2022, live, stream, click, watch, live] | [warner, robins, coffee, georgia, high, school, baseball, 2022, live, stream, click, watch, live] |
| **32** | [high, school, janet] | [witnessed, firsthand, high] |
| **33** | [memba, nigga, high, school] | [witnessed, firsthand, high] |
| **34** | [high, school, flight, school] | [witnessed, firsthand, high] |
| **35** | [high, school, janet] | [upgrading, high, school, theatre, lighting] |
| **36** | [memba, nigga, high, school] | [upgrading, high, school, theatre, lighting] |
| **37** | [high, school, flight, school] | [upgrading, high, school, theatre, lighting] |
| **38** | [witnessed, firsthand, high] | [upgrading, high, school, theatre, lighting] |
| **39** | [swansea, high, school, softball, live, live, swansea, varsity, softball, team, home, conference, game, today] | [buren, rogers, heritage, high, school, softball, live, live, rogers, heritage, varsity, softball, team, home, conference, game, buren, today] |
| **40** | [grand, magus, bellgrove, university] | [university] |
| **41** | [university, santo, tomas] | [university] |
| **42** | [university] | [university] |
| **43** | [university] | [university] |
| **44** | [bake, university] | [university] |
| **45** | [university, cease, exist] | [university] |
| **46** | [agile, university, university] | [university] |
| **47** | [warner, robins, coffee, georgia, high, school, baseball, 2022, live, stream, click, watch, live] | [north, springs, chamblee, georgia, high, school, baseball, 2022, live, stream, click, watch, live] |
| **48** | [high, school, baseball, live, genevieve, springs, yucca, live] | [high, school, baseball, live, blanchester, live] |
| **49** | [high, school, janet] | [high, school] |
| **50** | [memba, nigga, high, school] | [high, school] |
| **51** | [high, school, flight, school] | [high, school] |
| **52** | [witnessed, firsthand, high] | [high, school] |
| **53** | [upgrading, high, school, theatre, lighting] | [high, school] |
| **54** | [college, teammates] | [roman, college, zombies] |
| **55** | [high, school, janet] | [high, school, equivalency, karim] |
| **56** | [memba, nigga, high, school] | [high, school, equivalency, karim] |
| **57** | [high, school, flight, school] | [high, school, equivalency, karim] |
| **58** | [witnessed, firsthand, high] | [high, school, equivalency, karim] |
| **59** | [upgrading, high, school, theatre, lighting] | [high, school, equivalency, karim] |
| **60** | [high, school] | [high, school, equivalency, karim] |
| **61** | [thaw, grade, student, basic, education, high, school, fallen, defensing, genocidal, military, april] | [thaw, grade, student, basic, education, high, school, fallen, defensing, genocidal, military, april] |

| | | |
|---|---|---|
| 62 | [memorial, fpca, georgia, high, school, baseball, 2022, live, stream, click, watch, live] | [arcola, martinsville, illinois, high, school, baseball, 2022, live, stream, click, watch, live] |
| 63 | [ambridge, castle, pennsylvania, high, school, baseball, 2022, live, stream, click, watch, live] | [arcola, martinsville, illinois, high, school, baseball, 2022, live, stream, click, watch, live] |
| 64 | [warner, robins, coffee, georgia, high, school, baseball, 2022, live, stream, click, watch, live] | [arcola, martinsville, illinois, high, school, baseball, 2022, live, stream, click, watch, live] |
| 65 | [allatoona, timber, creek, colorado, high, school, baseball, 2022, live, stream, click, watch, live] | [arcola, martinsville, illinois, high, school, baseball, 2022, live, stream, click, watch, live] |
| 66 | [bellflower, anthony, california, high, school, baseball, 2022, live, stream, click, watch, live] | [arcola, martinsville, illinois, high, school, baseball, 2022, live, stream, click, watch, live] |
| 67 | [research, group, university, michigan, inviting, take, survey, social, media, survey, takes, receive, read, project, take, survey] | [inviting, take, survey, social, media, project, team, university, michigan, read, take, survey, survey, takes, receive] |
| 68 | [bellflower, anthony, california, high, school, baseball, 2022, live, stream, click, watch, live] | [southwestern, warren, east, kentucky, high, school, baseball, 2022, live, stream, click, watch, live] |
| 69 | [arcola, martinsville, illinois, high, school, baseball, 2022, live, stream, click, watch, live] | [southwestern, warren, east, kentucky, high, school, baseball, 2022, live, stream, click, watch, live] |
| 70 | [warner, robins, coffee, georgia, high, school, baseball, 2022, live, stream, click, watch, live] | [peck, ubly, michigan, high, school, baseball, 2022, live, stream, click, watch, live] |
| 71 | [arcola, martinsville, illinois, high, school, baseball, 2022, live, stream, click, watch, live] | [peck, ubly, michigan, high, school, baseball, 2022, live, stream, click, watch, live] |
| 72 | [high, school, baseball, live, indianola, live] | [high, school, baseball, live, robinson, worth, boca, live] |
| 73 | [high, school, baseball, live, blanchester, live] | [high, school, baseball, live, robinson, worth, boca, live] |
| 74 | [high, school, baseball, live, dover, bermudian, manheim, dubois, central, live] | [high, school, baseball, live, robinson, worth, boca, live] |
| 75 | [baseball, junior, pendleton, heights, high, school] | [junior, varsity, baseball, high, school, cullman, high, 2022, high, school, baseball, varsity, boys, varsity, boys, baseball] |
| 76 | [college, teammates] | [college, vijayawada] |
| 77 | [roman, college, zombies] | [college, vijayawada] |

In [4]:

```
import pandas as pd
import matplotlib.pyplot as plt
from pyspark.sql.functions import *
import re
from pyspark.ml.feature import MinHashLSH
from pyspark.ml.feature import CountVectorizer,  IDF, CountVectorizerModel, Tokenizer, RegexTokenizer, StopWordsR
sample = eng.limit(100000)
vectorize = CountVectorizer(inputCol="tokens", outputCol="features", minDF=1.0)
text_vectorize = vectorize.fit(sample).transform(sample)
mh = MinHashLSH(inputCol="features", outputCol="hashes", numHashTables=5)
text_model = mh.fit(text_vectorize)
text_hashed = mh.fit(text_vectorize).transform(text_vectorize)
jaccard_distance = 0.20

df_dups_text = text_model.approxSimilarityJoin(text_hashed, text_hashed, jaccard_distance).\
                filter("datasetA.id < datasetB.id").select(col("distCol"),\
             col("datasetA.id").alias("id_A"),
            col("datasetB.id").alias("id_B")
           )
```

In [ ]:

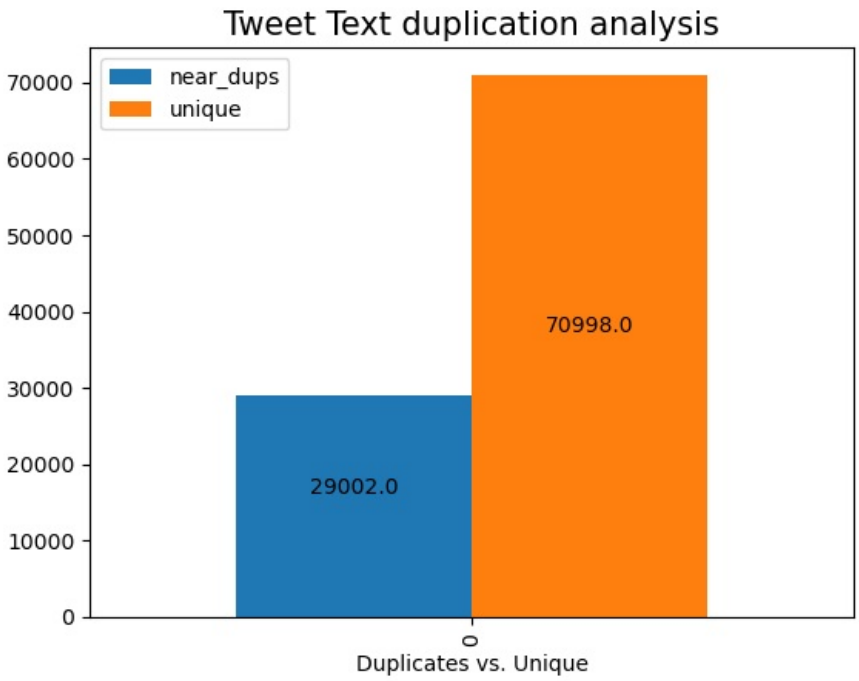
```
records = text_hashed.count()
dups = df_dups_text.select('id_A').distinct().count()
dups_50 = dups
uniques = records - dups

print ('Total records: ', records)
print ('Duplicate text based on {', jaccard_distance, '} jaccard distance: ', dups)
print ('Unique text based on {', jaccard_distance, '} jaccard distance: ', jaccard_distance, ': ', uniques)
```

```
22/12/05 04:23:07 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'.
22/12/05 04:35:35 WARN org.apache.spark.deploy.yarn.YarnAllocator: Container from a bad node: container_166969513
9548_0021_01_000008 on host: hub-msca-bdp-dphub-students-chenfeng-sw-zpsr.c.msca-bdp-students.internal. Exit stat
us: 143. Diagnostics: [2022-12-05 04:35:35.099]Container killed on request. Exit code is 143
[2022-12-05 04:35:35.099]Container exited with a non-zero exit code 143.
[2022-12-05 04:35:35.104]Killed by external signal
.
22/12/05 04:35:35 WARN org.apache.spark.scheduler.cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Requesting
driver to remove executor 8 for reason Container from a bad node: container_1669695139548_0021_01_000008 on host:
hub-msca-bdp-dphub-students-chenfeng-sw-zpsr.c.msca-bdp-students.internal. Exit status: 143. Diagnostics: [2022-1
2-05 04:35:35.099]Container killed on request. Exit code is 143
[2022-12-05 04:35:35.099]Container exited with a non-zero exit code 143.
```

```
[2022-12-05 04:35:35.104]Killed by external signal
.
22/12/05 04:35:35 ERROR org.apache.spark.scheduler.cluster.YarnScheduler: Lost executor 8 on hub-msca-bdp-dphub-students-chenfeng-sw-zpsr.c.msca-bdp-students.internal: Container from a bad node: container_1669695139548_0021_01_000008 on host: hub-msca-bdp-dphub-students-chenfeng-sw-zpsr.c.msca-bdp-students.internal. Exit status: 143. Diagnostics: [2022-12-05 04:35:35.099]Container killed on request. Exit code is 143
[2022-12-05 04:35:35.099]Container exited with a non-zero exit code 143.
[2022-12-05 04:35:35.104]Killed by external signal
.
22/12/05 04:35:35 WARN org.apache.spark.scheduler.TaskSetManager: Lost task 0.0 in stage 22.0 (TID 277) (hub-msca-bdp-dphub-students-chenfeng-sw-zpsr.c.msca-bdp-students.internal executor 8): ExecutorLostFailure (executor 8 exited caused by one of the running tasks) Reason: Container from a bad node: container_1669695139548_0021_01_000008 on host: hub-msca-bdp-dphub-students-chenfeng-sw-zpsr.c.msca-bdp-students.internal. Exit status: 143. Diagnostics: [2022-12-05 04:35:35.099]Container killed on request. Exit code is 143
[2022-12-05 04:35:35.099]Container exited with a non-zero exit code 143.
[2022-12-05 04:35:35.104]Killed by external signal
.
[Stage 22:>                                                         (0 + 1) / 1]
Total records:  100000
Duplicate text based on { 0.2 } jaccard distance:  29002
Unique text based on { 0.2 } jaccard distance:  0.2 :  70998
```

In [6]:

```
dups_df = pd.DataFrame.from_dict({'near_dups': [dups], 'unique': [uniques]})

ax=dups_df.plot(kind = 'bar',y=['near_dups', 'unique'], fontsize=10, color=['C0', 'C1'], align='center', width=0.
                xlabel="Duplicates vs. Unique")
ax.set_title('Tweet Text duplication analysis', fontsize=15)
for p in ax.patches:
       ax.annotate(format(p.get_height(), '.1f'),
                   (p.get_x() + p.get_width() / 2., p.get_height()/2),
                   ha = 'center', va = 'center',
                   xytext = (0, 9),
                   textcoords = 'offset points')
```




In [ ]:

```
spark.stop()
```

In [ ]:

```
spark = SparkSession.builder.config('spark.driver.maxResultSize','8g').config("spark.sql.shuffle.partitions",200)
config("spark.sql.broadcastTimeout", "36000").\
getOrCreate()
```

In [ ]:

```
print(spark./gateway/default/node/conf?host&port.get('spark.driver.maxResultSize'))
```

In [1]:

```
topic = spark.read.parquet('gs://'+'msca-bdp-students-bucket/shared_data/chenfeng/project/topic/')
```

In [78]:
```
topic.count()
```

Out[78]: 109509

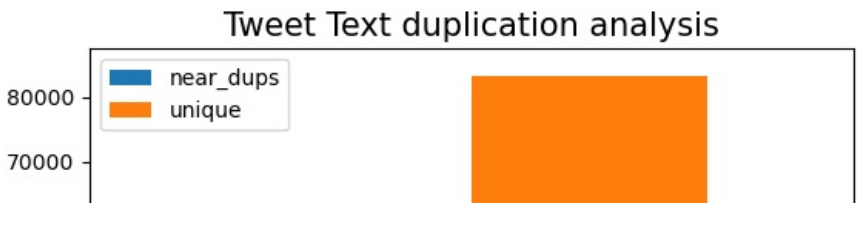
In [79]:
```
sample = topic.limit(100000)
```

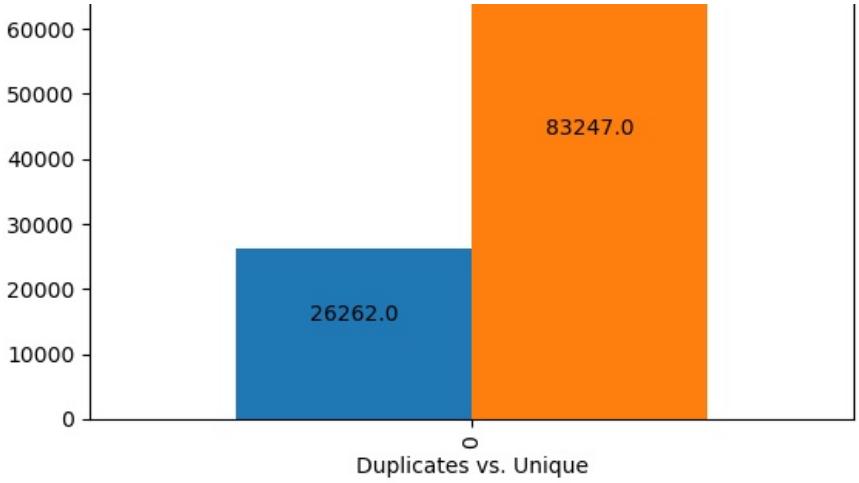
In [3]:
```
import pandas as pd
import matplotlib.pyplot as plt
from pyspark.sql.functions import *
import re
from pyspark.ml.feature import MinHashLSH
from pyspark.ml.feature import CountVectorizer,  IDF, CountVectorizerModel, Tokenizer, RegexTokenizer, StopWordsR
vectorize = CountVectorizer(inputCol="tokens", outputCol="features", minDF=1.0)
text_vectorize = vectorize.fit(topic).transform(topic)
mh = MinHashLSH(inputCol="features", outputCol="hashes", numHashTables=5)
text_model = mh.fit(text_vectorize)
text_hashed = mh.fit(text_vectorize).transform(text_vectorize)

jaccard_distance = 0.20

df_dups_text = text_model.approxSimilarityJoin(text_hashed, text_hashed, jaccard_distance).\
                filter("datasetA.id < datasetB.id").select(col("distCol"),\
             col("datasetA.id").alias("id_A"),
            col("datasetB.id").alias("id_B")
           )
```

In [ ]:
```
records = text_hashed.count()
dups = df_dups_text.select('id_A').distinct().count()
dups_50 = dups
uniques = records - dups

print ('Total records: ', records)
print ('Duplicate text based on {', jaccard_distance, '} jaccard distance: ', dups)
print ('Unique text based on {', jaccard_distance, '} jaccard distance: ', jaccard_distance, ': ', uniques)
```

```
[Stage 256:>                                                        (0 + 1) / 1]

Total records:  109509
Duplicate text based on { 0.2 } jaccard distance:  26262
Unique text based on { 0.2 } jaccard distance:  0.2 :  83247
```

In [82]:
```
dups_df = pd.DataFrame.from_dict({'near_dups': [dups], 'unique': [uniques]})

ax=dups_df.plot(kind = 'bar',y=['near_dups', 'unique'], fontsize=10, color=['C0', 'C1'], align='center', width=0.
                xlabel="Duplicates vs. Unique")
ax.set_title('Tweet Text duplication analysis', fontsize=15)
for p in ax.patches:
       ax.annotate(format(p.get_height(), '.1f'),
                   (p.get_x() + p.get_width() / 2., p.get_height()/2),
                   ha = 'center', va = 'center',
                   xytext = (0, 9),
                   textcoords = 'offset points')
```






In [ ]:

```
d = df_dups_text.limit(10000)
d.show()
```

```
[Stage 273:>                                                        (0 + 1) / 1]
```

```
+------------------+-------------------+-------------------+
|           distCol|               id_A|               id_B|
+------------------+-------------------+-------------------+
|               0.0|1511207384907145220|1511301419793534976|
|               0.0|1511207384907145220|1511531710714105858|
|               0.0|1511207384907145220|1511576962002788355|
|               0.0|1511207384907145220|1511660484105478148|
|               0.0|1511207384907145220|1511715364618764289|
|               0.0|1511207384907145220|1512403743412989955|
|               0.0|1511207384907145220|1512665368938299393|
|               0.0|1511207384907145220|1512695796348878849|
|               0.0|1511207384907145220|1512774842130518020|
|               0.0|1511207384907145220|1513377568099086339|
|               0.0|1511207384907145220|1513421749462843394|
|               0.0|1511207384907145220|1513736183490166788|
|               0.0|1511207384907145220|1513879625180803074|
|0.1428571428571429|1511207384907145220|1514786151571533824|
|0.1428571428571429|1511207384907145220|1514834762791530496|
|0.1428571428571429|1511207384907145220|1514878781487738882|
|0.1428571428571429|1511207384907145220|1514970589186113538|
|0.1428571428571429|1511207384907145220|1515171068402868226|
|0.1428571428571429|1511207384907145220|1515196066190471173|
|0.1428571428571429|1511207384907145220|1515233591462359042|
+------------------+-------------------+-------------------+
only showing top 20 rows
```

In [6]:

```
text = spark.read.parquet('gs://'+'msca-bdp-students-bucket/shared_data/chenfeng/project/key_eda/')
original = text.filter(col('retweet').isNull())
ori_text = original.select('id','full_text')
df = df_dups_text.limit(10000).join(ori_text, df_dups_text.id_A == ori_text.id, 'inner').select(df_dups_text.id_A
df = df.withColumnRenamed('full_text', 'full_text_A')
df = df.join(ori_text, df_dups_text.id_B == ori_text.id, 'inner').drop(ori_text.id)
df = df.withColumnRenamed('full_text', 'full_text_B')
df.show()
```

```
22/12/06 15:10:26 WARN org.apache.spark.deploy.yarn.YarnAllocator: Container from a bad node: container_167012515
3499_0010_01_000005 on host: hub-msca-bdp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal. Ex
it status: 143. Diagnostics: [2022-12-06 15:10:26.472]Container killed on request. Exit code is 143
[2022-12-06 15:10:26.473]Container exited with a non-zero exit code 143.
[2022-12-06 15:10:26.486]Killed by external signal
.
22/12/06 15:10:26 WARN org.apache.spark.deploy.yarn.YarnAllocator: Container from a bad node: container_167012515
3499_0010_01_000006 on host: hub-msca-bdp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal. Ex
it status: 143. Diagnostics: [2022-12-06 15:10:26.508]Container killed on request. Exit code is 143
[2022-12-06 15:10:26.508]Container exited with a non-zero exit code 143.
[2022-12-06 15:10:26.509]Killed by external signal
.
22/12/06 15:10:26 WARN org.apache.spark.scheduler.cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Requesting
driver to remove executor 4 for reason Container from a bad node: container_1670125153499_0010_01_000005 on host:
```

```
hub-msca-bdp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal. Exit status: 143. Diagnostics:
[2022-12-06 15:10:26.472]Container killed on request. Exit code is 143
[2022-12-06 15:10:26.473]Container exited with a non-zero exit code 143.
[2022-12-06 15:10:26.486]Killed by external signal
.
22/12/06 15:10:26 WARN org.apache.spark.scheduler.cluster.YarnSchedulerBackend$YarnSchedulerEndpoint: Requesting
driver to remove executor 5 for reason Container from a bad node: container_1670125153499_0010_01_000006 on host:
hub-msca-bdp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal. Exit status: 143. Diagnostics:
[2022-12-06 15:10:26.508]Container killed on request. Exit code is 143
[2022-12-06 15:10:26.508]Container exited with a non-zero exit code 143.
[2022-12-06 15:10:26.509]Killed by external signal
.
22/12/06 15:10:26 ERROR org.apache.spark.scheduler.cluster.YarnScheduler: Lost executor 5 on hub-msca-bdp-dphub-s
tudents-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal: Container from a bad node: container_1670125153499_
0010_01_000006 on host: hub-msca-bdp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal. Exit st
atus: 143. Diagnostics: [2022-12-06 15:10:26.508]Container killed on request. Exit code is 143
[2022-12-06 15:10:26.508]Container exited with a non-zero exit code 143.
[2022-12-06 15:10:26.509]Killed by external signal
.
22/12/06 15:10:26 WARN org.apache.spark.scheduler.TaskSetManager: Lost task 7.0 in stage 8.0 (TID 92) (hub-msca-b
dp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal executor 5): ExecutorLostFailure (executor
5 exited caused by one of the running tasks) Reason: Container from a bad node: container_1670125153499_0010_01_0
00006 on host: hub-msca-bdp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal. Exit status: 143
. Diagnostics: [2022-12-06 15:10:26.508]Container killed on request. Exit code is 143
[2022-12-06 15:10:26.508]Container exited with a non-zero exit code 143.
[2022-12-06 15:10:26.509]Killed by external signal
.
22/12/06 15:10:26 WARN org.apache.spark.scheduler.TaskSetManager: Lost task 1.0 in stage 8.0 (TID 86) (hub-msca-b
dp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal executor 5): ExecutorLostFailure (executor
5 exited caused by one of the running tasks) Reason: Container from a bad node: container_1670125153499_0010_01_0
00006 on host: hub-msca-bdp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal. Exit status: 143
. Diagnostics: [2022-12-06 15:10:26.508]Container killed on request. Exit code is 143
[2022-12-06 15:10:26.508]Container exited with a non-zero exit code 143.
[2022-12-06 15:10:26.509]Killed by external signal
.
22/12/06 15:10:26 ERROR org.apache.spark.scheduler.cluster.YarnScheduler: Lost executor 4 on hub-msca-bdp-dphub-s
tudents-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal: Container from a bad node: container_1670125153499_
0010_01_000005 on host: hub-msca-bdp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal. Exit st
atus: 143. Diagnostics: [2022-12-06 15:10:26.472]Container killed on request. Exit code is 143
[2022-12-06 15:10:26.473]Container exited with a non-zero exit code 143.
[2022-12-06 15:10:26.486]Killed by external signal
.
22/12/06 15:10:26 WARN org.apache.spark.scheduler.TaskSetManager: Lost task 4.0 in stage 8.0 (TID 89) (hub-msca-b
dp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal executor 4): ExecutorLostFailure (executor
4 exited caused by one of the running tasks) Reason: Container from a bad node: container_1670125153499_0010_01_0
00005 on host: hub-msca-bdp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal. Exit status: 143
. Diagnostics: [2022-12-06 15:10:26.472]Container killed on request. Exit code is 143
[2022-12-06 15:10:26.473]Container exited with a non-zero exit code 143.
[2022-12-06 15:10:26.486]Killed by external signal
.
22/12/06 15:10:26 WARN org.apache.spark.scheduler.TaskSetManager: Lost task 8.0 in stage 8.0 (TID 93) (hub-msca-b
dp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal executor 4): ExecutorLostFailure (executor
4 exited caused by one of the running tasks) Reason: Container from a bad node: container_1670125153499_0010_01_0
00005 on host: hub-msca-bdp-dphub-students-backup-chenfeng-sw-wlgd.c.msca-bdp-students.internal. Exit status: 143
. Diagnostics: [2022-12-06 15:10:26.472]Container killed on request. Exit code is 143
[2022-12-06 15:10:26.473]Container exited with a non-zero exit code 143.
[2022-12-06 15:10:26.486]Killed by external signal
.
```

| id_A | full_text_A | id_B | full_text_B |
|---|---|---|---|
| 1512166406498156545 | Kathleen vs All S... | 1513982956033175553 | King's Academy vs... |
| 1517135625111367680 | Delhi University ... | 1517144255126327300 | Delhi University ... |
| 1517139155179900929 | Delhi University ... | 1517144255126327300 | Delhi University ... |
| 1526431941024796673 | Weather summary f... | 1528972958340898816 | Weather summary f... |
| 1527152334970802177 | Weather summary f... | 1528972958340898816 | Weather summary f... |
| 1528546281110528000 | Sunset Timelapse ... | 1531446978042904576 | Sunset Timelapse ... |
| 1535222564246757378 | Sunrise Timelapse... | 1538122025964867585 | Sunrise Timelapse... |
| 1513528133223333898 | Many students com... | 1539786136721264647 | Many students com... |
| 1531141677695619074 | Many students com... | 1539786136721264647 | Many students com... |
| 1511565183159840769 | Many students com... | 1539786136721264647 | Many students com... |
| 1524053509364142080 | Just posted a pho... | 1540168186104094723 | Just posted a pho... |
| 1534666228022771716 | Just posted a pho... | 1540168186104094723 | Just posted a pho... |
| 1538078489370411009 | @nduuh_wheels @Mv... | 1543690920486486016 | @Luntu_Kheswa Is ... |
| 1545644432225091585 | My mum bought me ... | 1547063250126524416 | My mum bought me ... |
| 1530508077472137216 | Lifestyle forecas... | 1547181481075908609 | Lifestyle forecas... |
| 1533044839616393216 | Lifestyle forecas... | 1547181481075908609 | Lifestyle forecas... |
| 1543557809198997507 | Lifestyle forecas... | 1547181481075908609 | Lifestyle forecas... |
| 1515302781514653699 | Lifestyle forecas... | 1547181481075908609 | Lifestyle forecas... |
| 1520365193976598532 | Lifestyle forecas... | 1548631130865340425 | Lifestyle forecas... |
| 1515291750323134464 | Lifestyle forecas... | 1548631130865340425 | Lifestyle forecas... |

only showing top 20 rows

In [57]:
```
pd.set_option('display.max_rows', 30)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('max_colwidth', -1)
df_pd[['full_text_A','full_text_B']]
```

```
/tmp/ipykernel_30821/1424795411.py:3: FutureWarning: Passing a negative integer is deprecated in version 1.0 and
will not be supported in future version. Instead, use None to not limit the column width.
  pd.set_option('max_colwidth', -1)
```

Out[57]:

| | full_text_A | full_text_B |
|---|---|---|
| **0** | @LittleToller Hi Gracie,\nHere are some details to book tickets for an event being held at the London South Bank University on the 25th of November.\nhttps://t.co/MoniLc2aWj\nI hoe you can make this event\nBestwishes\n@lloyd_page | @LdSentongo Hi Jenny,\nHere are some details to book tickets for an event being held at the London South Bank University on the 25th of November.\nhttps://t.co/MoniLbL7Uj\nBest wishes\n@lloyd_page |
| **1** | @LittleToller Hi Gracie,\nHere are some details to book tickets for an event being held at the London South Bank University on the 25th of November.\nhttps://t.co/MoniLc2aWj\nI hoe you can make this event\nBestwishes\n@lloyd_page | @SueBprof1 Hi Sue,\nHere are some details to book tickets for an event being held at the London South Bank University on the 25th of November.\nhttps://t.co/MoniLbL7Uj\nBest wishes\n@lloyd_page |
| **2** | @LittleToller Hi Gracie,\nHere are some details to book tickets for an event being held at the London South Bank University on the 25th of November.\nhttps://t.co/MoniLc2aWj\nI hoe you can make this event\nBestwishes\n@lloyd_page | @jimgblair Hi Jim,\nHere are some details to book tickets for an event being held at the London South Bank University on the 25th of November.\nhttps://t.co/MoniLc2aWj\nBest wishes\n@lloyd_page |
| **3** | @LittleToller Hi Gracie,\nHere are some details to book tickets for an event being held at the London South Bank University on the 25th of November.\nhttps://t.co/MoniLc2aWj\nI hoe you can make this event\nBestwishes\n@lloyd_page | @OxBugBuster Hi Jaz,\nHere are some details to book tickets for an event being held at the London South Bank University on the 25th of November.\nhttps://t.co/MoniLbL7Uj\nBest wishes\n@lloyd_page |
| **4** | @LittleToller Hi Gracie,\nHere are some details to book tickets for an event being held at the London South Bank University on the 25th of November.\nhttps://t.co/MoniLc2aWj\nI hoe you can make this event\nBestwishes\n@lloyd_page | @Daniel_Watkins1 Hi Daniel,\nHere are some details to book tickets for an event being held at the London South Bank University on the 25th of November.\nhttps://t.co/MoniLc2aWj\nBest wishes\n@lloyd_page |
| **...** | ... | ... |
| **32060** | Florida State University Duffel Bag - FSU Gym Bags w/ SHOE POCKET [0CSRCC9]\n\nhttps://t.co/Mbi2BfzTdv | Florida State University Duffel Bag - FSU Gym Bags w/ SHOE POCKET [RXFNWGM]\n\nhttps://t.co/ywh7YSnTSF |
| **32061** | My high school math teacher would def say this | just saw my high school math teacher on the tl |
| **32062** | My high school math teacher would def say this | me as a high school math teacher |
| **32063** | My high school math teacher would def say this | @MockDraftMonday His mom was my high school math teacher lol |
| **32064** | My high school math teacher would def say this | @NutmegStOfMind Your high school math teacher would be so proud of you |

32065 rows × 2 columns

In [81]:
```
duplicate = spark.read.parquet('gs://'+'msca-bdp-students-bucket/shared_data/chenfeng/project/duplicate/')
text = spark.read.parquet('gs://'+'msca-bdp-students-bucket/shared_data/chenfeng/project/key_eda/')
original = text.filter(col('retweet').isNull())
```

In [11]:
```
duplicate.show()
```

```
+------------------+--------------------+------------------+--------------------+
|              id_A|          full_text_A|              id_B|          full_text_B|
+------------------+--------------------+------------------+--------------------+
|1513157074456559621|If a book contain...|1516101463596224512|If a book contain...|
|1516631011274145795|Weather summary f...|1516643032719888389|Weather summary f...|
|1513838741827301385|Lifestyle forecas...|1518190696225779715|Lifestyle forecas...|
|1520188526821883905|Florida Gulf Coas...|1522743694012686339|Sunset Timelapse ...|
|1523669604857712648|@unlvfootball Whe...|1524837415336169495|@EVNavyFB When y'...|
|1524905648043053065|Florida A&amp;M U...|1525265305378902016|University of Flo...|
|1522388308327813129|Sunset Timelapse ...|1531071640800477186|University of Wes...|
|1524962135838801920|Florida High Scho...|1531966539485720576|Florida High Scho...|
|1533063550389964800|Daily almanac for...|1536325041897590785|Daily almanac for...|
|1538316283246256128|BCEM Florida Gate...|1540491053475807232|BCEM Florida Gate...|
|1516219167879450628|Sunset Timelapse ...|1541956475152736260|Sunset Timelapse ...|
|1528909127396864002|Sunset Timelapse ...|1541956475152736260|Sunset Timelapse ...|
|1513318312691879937|Sunset Timelapse ...|1541956475152736260|Sunset Timelapse ...|
|1533044839616393216|Lifestyle forecas...|1542108706460127234|Lifestyle forecas...|
|1541340640918790145|@LivyMat17 @diteb...|1543696388273442817|@Fact Is there AN...|
|1536662950173454336|@lazyirv @HowsYou...|1543696388273442817|@Fact Is there AN...|
|1541338835522568193|@Naturallyamdope ...|1543696388273442817|@Fact Is there AN...|
|1538076226941755392|@AdvoBarryRoux Is...|1545655228703424512|@kaitareja Is the...|
```

```
|1538078699664416768|@Nqubeko__Shandu ...|1547180234654507010|@nmzima19 @tebogo...|
|1519120420703772672|Sunset Timelapse ...|1549203709573435393|Sunset Timelapse ...|
+-------------------+--------------------+-------------------+--------------------+
only showing top 20 rows
```

In [82]:
```
dup_user = duplicate.join(original, duplicate.id_A == original.id, 'inner').select('user_name','user_descrip','us
                                                                                    duplicate.id_A, 'full_text_A','id_B','ful
```

In [89]:
```
dup_user.show()
```

```
[Stage 173:=================================================>  (19 + 1) / 20]
+---------------+--------------------+-------------------+-------------------+--------------------+-------------
------+--------------------+
|      user_name|        user_descrip|            user_id|               id_A|         full_text_A|
id_B|         full_text_B|
+---------------+--------------------+-------------------+-------------------+--------------------+-------------
------+--------------------+
|      UWFWxSTEM|Providing real-ti...|         3135886028|1512756994435862528|University of Wes...|15616767880525
12770|University of Wes...|
|      UWFWxSTEM|Providing real-ti...|         3135886028|1512756994435862528|University of Wes...|15330401471844
67968|University of Wes...|
|      UWFWxSTEM|Providing real-ti...|         3135886028|1512756994435862528|University of Wes...|15660278300682
19904|University of Wes...|
|      UWFWxSTEM|Providing real-ti...|         3135886028|1512756994435862528|University of Wes...|15769033387922
06337|University of Wes...|
|      UWFWxSTEM|Providing real-ti...|         3135886028|1512756994435862528|University of Wes...|15413755329605
67299|University of Wes...|
|  RealChanakya_|no necesita te a ...|1264409578092937222|1514855874061209601|@msisodia #DUStud...|15148562462805
40161|@msisodia #DUStud...|
|    dailyseung2|— ☆ for stray kid...|1446441613723533315|1514967612941549577|@namudwaekki i lo...|15604843269863
01441|@ChrisWalkerCBC I...|
|      UCFWxSTEM|Providing real-ti...|         3308132768|1516010552694394886|University of Cen...|15736358591312
77312|University of Cen...|
|      UCFWxSTEM|Providing real-ti...|         3308132768|1516010552694394886|University of Cen...|15232545068883
51745|University of Cen...|
|      UCFWxSTEM|Providing real-ti...|         3308132768|1516010552694394886|University of Cen...|15569596694359
98208|University of Cen...|
|      UCFWxSTEM|Providing real-ti...|         3308132768|1516010552694394886|University of Cen...|15605847110927
68768|University of Cen...|
|      UCFWxSTEM|Providing real-ti...|         3308132768|1516010552694394886|University of Cen...|15613098093743
71840|University of Cen...|
|      UCFWxSTEM|Providing real-ti...|         3308132768|1516010552694394886|University of Cen...|15163726475417
10858|University of Cen...|
|student76737126|    Delhi University|1515763788087459841|1517151058229084160|Delhi University ...|15175117396337
17248|@ThePradeepRawat ...|
|student76737126|    Delhi University|1515763788087459841|1517151058229084160|Delhi University ...|15175116224238
91969|Delhi University ...|
|student76737126|    Delhi University|1515763788087459841|1517151058229084160|Delhi University ...|15173867451592
37632|@ThePradeepRawat ...|
|LubzStaProperty|Cabinet Member Pa...|          249452687|1521570660883083266|Platinum Jubilee ...|15715031139111
28064|Platinum Jubilee ...|
|     FPTCWxSTEM|Providing real-ti...|         3286764350|1525270330817781760|Florida Panhandle...|15850428467547
66850|Florida Panhandle...|
|     FPTCWxSTEM|Providing real-ti...|         3286764350|1525270330817781760|Florida Panhandle...|15404951399315
61985|Florida Panhandle...|
|     FPTCWxSTEM|Providing real-ti...|         3286764350|1525270330817781760|Florida Panhandle...|15317954114967
51111|Florida Panhandle...|
+---------------+--------------------+-------------------+-------------------+--------------------+-------------
------+--------------------+
only showing top 20 rows
```

In [83]:
```
import re
import nltk
nltk.download('stopwords')
from nltk.corpus import stopwords
dup_user = dup_user.filter(col('user_descrip').isNotNull())
d = dup_user.rdd.map(lambda x : x['user_descrip']).filter(lambda x: x is not None)
StopWords = stopwords.words("english")
# remove stop words
tokens = d\
```

```
    .map( lambda document: document.strip().lower())\
    .map( lambda document: re.sub("@[A-Za-z0-9_]+","", document))\
    .map( lambda document: re.sub(r'[^\w\s]', '', document))\
    .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])
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

In [84]:
```
t = tokens.zip(dup_user.select('id_A').rdd.flatMap(lambda x:x))
df_tokens = t.toDF(['user_tokens','id_A'])
```

In [85]:
```
df_tokens.write.saveAsTable('d', mode = 'overwrite')
```

In [213…
```
query = """
        select * from d where array_contains (user_tokens, 'sports') \
        """
```

In [214…
```
df = spark.sql(query)
```

In [215…
```
df.count()
```

Out[215… 57

In [230…
```
df.limit(20).show()
```

```
+--------------------+------------------+
|         user_tokens|              id_A|
+--------------------+------------------+
|[wdrb, news, affi...|1549008171364225025|
|[husband, avid, s...|1585638911535382529|
|[professional, ph...|1549802839043723265|
|[published, twice...|1520026131025612802|
|[published, twice...|1520026131025612802|
|[published, twice...|1520026131025612802|
|[professional, ph...|1549395908479049731|
|[professor, nutri...|1536319172849938432|
|[professor, nutri...|1536319172849938432|
|[make, smarter, s...|1575733638234222592|
|[alumna, senior, ...|1524053509364142080|
|[alumna, senior, ...|1524053509364142080|
|[alumna, senior, ...|1524053509364142080|
|[alumna, senior, ...|1524053509364142080|
|[alumna, senior, ...|1524053509364142080|
|[alumna, senior, ...|1524053509364142080|
|[alumna, senior, ...|1524053509364142080|
|[alumna, senior, ...|1524053509364142080|
|[alumna, senior, ...|1524053509364142080|
+--------------------+------------------+
```

## Politics Entites Tweet Duplication

In [130…
```
dup_user.filter(col('id_A')==1570024960646594560).limit(1).collect()
```

Out[130…
```
[Row(user_name='PKFLRDA', user_descrip='Retired; Ph.D.Engg.;Global Bus. Mgmt Exp;Life Master-Bridge; Golf;Reading
; Computers; Hindi, English Pop, and Oriya songs; Bollywood Videos; USA Politics & News', user_id=88427831, id_A=
1570024960646594560, full_text_A="'What is woke math?': In Florida, public school teachers bristle at DeSantis's
changes to education https://t.co/cPyLOY6rJp via @YahooNews - @AP @NicolleDWallace @LeaderMcConnell @GOPLeader @S
enSchumer @SpeakerPelosi #Florida @GovRonDeSantis @CharlieCrist", id_B=1570262311503540224, full_text_B="'What is
woke math?': In Florida, public school teachers bristle at DeSantis's changes to education https://t.co/viMvu8nK7
4")]
```

In [133…
```
dup_user.filter(col('id_A')==1585764447611543553).collect()
```

Out[133…
```
[Row(user_name='PapaESoCo', user_descrip="Nothing to see here: Old Guy that express's his opinion on Politics and
Scoundrels; they go together so well! #BlackLivesMatter #SFGIANTS #BidenHarris2020", user_id=2173135652, id_A=158
5764447611543553, full_text_A="'No confidence' in Ben Sasse: Florida college gives senator the thumbs down https:
//t.co/o38kLnaZdU Good News!", id_B=1585778407211765760, full_text_B="'No confidence' in Ben Sasse: Florida colle
ge gives senator the thumbs down #SmartNews  https://t.co/37BrvLGqPy")]
```

In [136…
```
dup_user.filter(col('id_A')==1570191966725615616).collect()
```

Out[136…
```
[Row(user_name='CLeinwand12', user_descrip='Enjoy Dachshunds & politics - Dogs must look at people and wonder, wh
o put them in charge?', user_id=22699560, id_A=1570191966725615616, full_text_A="'What is woke math?': In Florida
, public school teachers bristle at DeSantis's changes to education https://t.co/wQEipRhrBN", id_B=15702149602986
10688, full_text_B="'What is woke math?': In Florida, public school teachers bristle at DeSantis's changes to edu
cation https://t.co/PZuRWpobsR via @Yahoo")]
```

In [201…
```
dup_user.filter(col('id_A')==1587579179138256896).collect()
```

Out[201…
```
[Row(user_name='VeraldoF4F', user_descrip='CONSERVATIVE LOVE MY FAMILY, MY COUNTRY TRUTH WILL SET YOU FREE  DM
DATING □SALES □IFB  □□ □□ #AMERICAFIRST #FREEAMERICA #SAVEAMERICA #MAGA #FJB', user_id=151304629, id_A=1587579179
8256896, full_text_A="81 MILLION VOTES!!!! Joe Biden Can't Fill Up Small College Gym at Florida Rally with Duds C
rist and Val Demings - (VIDEO) https://t.co/Do0NDypuJ5 via @gatewaypundit", id_B=1587944121897033728, full_text_B
="81 Million Votes! Joe Biden Can't Fill Up Small College Gym at Florida Rally with Duds Crist and Val Demings -
VIDEO https://t.co/vMp820veDB via @gatewaypundit")]
```

In [202…
```
dup_user.filter(col('id_A')==1579947705610010624).collect()
```

Out[202…
```
[Row(user_name='USAfree1979', user_descrip='Constitutional conservative, God,Wife,Country support 1A,2A pro-life'
, user_id=1521466693326753800, id_A=1579947705610010624, full_text_A="University of Florida students protest, say
Republican Sen. Ben Sasse poses 'threat' as president\n\nhttps://t.co/GOqkKayhWk", id_B=1579954160346140672, full
_text_B="University of Florida students protest, say Republican Sen. Ben Sasse poses 'threat' as president\n\nhtt
ps://t.co/Ci0OgkQedb")]
```

In [233…
```
ls = [1570024960646594560, 1585764447611543553,1570191966725615616,1587579179138256896,1579947705610010624]
politics = dup_user.filter(dup_user.id_A.isin(ls)).limit(10).toPandas()
pd.set_option('display.max_rows', 30)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('max_colwidth', -1)
politics[['user_name','full_text_A']]
```

```
/tmp/ipykernel_17136/1918221796.py:3: FutureWarning: Passing a negative integer is deprecated in version 1.0 and
will not be supported in future version. Instead, use None to not limit the column width.
  pd.set_option('max_colwidth', -1)
```

Out[233…

| | user_name | full_text_A |
|---|---|---|
| 0 | USAfree1979 | University of Florida students protest, say Republican Sen. Ben Sasse poses 'threat' as president\n\nhttps://t.co/GOqkKayhWk |

| | | |
|---|---|---|
| 1 | PKFLRDA | 'What is woke math?': In Florida, public school teachers bristle at DeSantis's changes to education https://t.co/cPyLOY6rJp via @YahooNews - @AP @NicolleDWallace @LeaderMcConnell @GOPLeader @SenSchumer @SpeakerPelosi #Florida @GovRonDeSantis @CharlieCrist |
| 2 | PKFLRDA | 'What is woke math?': In Florida, public school teachers bristle at DeSantis's changes to education https://t.co/cPyLOY6rJp via @YahooNews - @AP @NicolleDWallace @LeaderMcConnell @GOPLeader @SenSchumer @SpeakerPelosi #Florida @GovRonDeSantis @CharlieCrist |
| 3 | PKFLRDA | 'What is woke math?': In Florida, public school teachers bristle at DeSantis's changes to education https://t.co/cPyLOY6rJp via @YahooNews - @AP @NicolleDWallace @LeaderMcConnell @GOPLeader @SenSchumer @SpeakerPelosi #Florida @GovRonDeSantis @CharlieCrist |
| 4 | PKFLRDA | 'What is woke math?': In Florida, public school teachers bristle at DeSantis's changes to education https://t.co/cPyLOY6rJp via @YahooNews - @AP @NicolleDWallace @LeaderMcConnell @GOPLeader @SenSchumer @SpeakerPelosi #Florida @GovRonDeSantis @CharlieCrist |
| 5 | PKFLRDA | 'What is woke math?': In Florida, public school teachers bristle at DeSantis's changes to education https://t.co/cPyLOY6rJp via @YahooNews - @AP @NicolleDWallace @LeaderMcConnell @GOPLeader @SenSchumer @SpeakerPelosi #Florida @GovRonDeSantis @CharlieCrist |
| 6 | VeraldoF4F | 81 MILLION VOTES!!!! Joe Biden Can't Fill Up Small College Gym at Florida Rally with Duds Crist and Val Demings - (VIDEO) https://t.co/Do0NDypuJ5 via @gatewaypundit |
| 7 | CLeinwand12 | 'What is woke math?': In Florida, public school teachers bristle at DeSantis's changes to education https://t.co/wQEipRhrBN |
| 8 | PapaESoCo | 'No confidence' in Ben Sasse: Florida college gives senator the thumbs down https://t.co/o38kLnaZdU Good News! |

## Health Org Tweet Duplication

In [141…
```
dup_user.filter(col('id_A')==1534428865363419137).limit(1).collect()
```

Out[141…
```
[Row(user_name='HRCH_NHS', user_descrip='Official feed for Hounslow and Richmond Community Healthcare NHS Trust. Tweets not monitored 24/7. Call 111 for urgent health advice and 999 in emergencies.', user_id=466463929, id_A=1534428865363419137, full_text_A='West Middlesex University Hospital Vaccination Hub\nPfizer vaccination clinic, 16 + year olds\nTuesday 7 and Friday 10 and Saturday 11 June, 8.30am to 3.20pm \nTo book, visit https://t.co/azm7PpyI3c https://t.co/Pf7BQSZNVv', id_B=1535243524022624259, full_text_B="West Middlesex University Hospital Vaccination Hub\nNext week's COVID vaccination clinics\nPfizer vaccination clinic, 16+ year olds\nTuesday 13 and Friday 14 and Saturday 17 June 2022, 8.30am to 3.20pm \nTo book, visit https://t.co/azm7PpyI3c https://t.co/ofPVYBKS9u")]
```

In [142…
```
dup_user.filter(col('id_A')==1529560467173167106).limit(1).collect()
```

Out[142…
```
[Row(user_name='DrSusanAnderso1', user_descrip='Lisc. Psychologist & Yoga Instructor specializing in “mindset conditioning” & mental health with athletes, health & human sexuality & trauma.', user_id=1365412810377863169, id_A=1529560467173167106, full_text_A="A Florida class president couldn't discuss being gay in high school graduation speech -- so he talked about his curly hair\nhttps://t.co/U4ZQ8sxwwp", id_B=1529823827231768576, full_text_B="A Florida class president couldn't discuss being gay in high school graduation speech -- so he talked about his curly hair\n#dontsaygay #florida #curlyhair\n\nhttps://t.co/QLYcHyb340")]
```

In [143…
```
dup_user.filter(col('id_A')==1533779295763632129).limit(1).collect()
```

Out[143…
```
[Row(user_name='JKPBooks', user_descrip='Books that make a difference on social care, mental health, education and more. @JKPAutism, @JKPGender, @JKPDementia & @JKPHealth', user_id=50276476, id_A=1533779295763632129, full_text_A='"A highly engaging book exploring relationships, #consent, respect, sexuality and identity in an accessible way...a coherent, inclusive and joined up approach to young people\'s development."\n\nProfessor Richard Woolley, University of Hull\n\n#RSE #SexualityEducation https://t.co/YXtGDoCGyX', id_B=1539939924660703235, full_text_B='"A highly engaging book exploring relationships, #consent, respect, sexuality and identity in an accessible way... a coherent, inclusive and joined up approach to young people\'s development."\n\nProfessor Richard Woolley, University of Hull\n\n#RSE #SexualityEducation https://t.co/uTYXFLzmRd')]
```

In [ ]:
```
### These tweets are not about florida book ban
```

In [234…
```
ls = [1534428865363419137, 1533779295763632129, 1529560467173167106]
health = dup_user.filter(dup_user.id_A.isin(ls)).limit(10).toPandas()
pd.set_option('display.max_rows', 30)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('max_colwidth', -1)
health[['user_name','full_text_A']]
```

```
/tmp/ipykernel_17136/2541917486.py:5: FutureWarning: Passing a negative integer is deprecated in version 1.0 and
will not be supported in future version. Instead, use None to not limit the column width.
  pd.set_option('max_colwidth', -1)
```

Out[234…

| | user_name | full_text_A |
|---|---|---|
| 0 | DrSusanAnderso1 | A Florida class president couldn't discuss being gay in high school graduation speech -- so he talked about his curly hair\nhttps://t.co/U4ZQ8sxwwp |
| 1 | JKPBooks | "A highly engaging book exploring relationships, #consent, respect, sexuality and identity in an accessible way...a coherent, inclusive and joined up approach to young people's development."\n\nProfessor Richard Woolley, University of Hull\n\n#RSE #SexualityEducation https://t.co/YXtGDoCGyX |
| 2 | HRCH_NHS | West Middlesex University Hospital Vaccination Hub\nPfizer vaccination clinic, 16+ year olds\nTuesday 7 and Friday 10 and Saturday 11 June, 8.30am to 3.20pm \nTo book, visit https://t.co/azm7PpyI3c https://t.co/Pf7BQSZNVv |

## News Outlets Tweet Duplication

In [148…

```
dup_user.filter(col('id_A')==1518064686620721155).limit(1).collect()
```

Out[148…

```
[Row(user_name='FoxNews', user_descrip="Follow America's #1 cable news network, delivering you breaking news, ins
ightful analysis, and must-see videos. http://foxnews.com/contact", user_id=1367531, id_A=1518064686620721155, fu
ll_text_A="University professors unhappy by Florida Gov. DeSantis' attempt to hold faculty 'accountable' https://
t.co/lHAUaoGaWO", id_B=1518362045791363072, full_text_B="University professors unhappy by Florida Gov. DeSantis'
attempt to hold faculty 'accountable'\n\nhttps://t.co/fl7taODIfJ")]
```

In [154…

```
dup_user.filter(col('id_A')==1528829779100635136).limit(1).collect()
```

Out[154…

```
[Row(user_name='BerkleyBearNews', user_descrip='Bringing the news that is important, that as quick as a dog can b
ring it.', user_id=787546010, id_A=1528829779100635136, full_text_A="Florida university must reinstate professor
who was fired over 'Black privilege' tweets https://t.co/wvrzMdMrFU #news #topstories #berkleybearnews", id_B=152
8929615254802433, full_text_B="Florida university must reinstate professor who was fired over 'Black privilege' t
weets | Fox News https://t.co/7y5wh3Lgzg")]
```

In [153…

```
dup_user.filter(col('id_A')==1512489448675876865).limit(1).collect()
```

Out[153…

```
[Row(user_name='ChannelRadio1', user_descrip='Channel Radio - an eclectic mix of top music shows, news, business,
chat, celebrity interviews, arts and drama. Also available via Tunein.', user_id=855209496, id_A=1512489448675876
865, full_text_A='The Kent and Medway Business Summit returns on Wednesday 27 April 2022 in the stunning Sibson b
uilding on the Canterbury campus of the University of Kent\n\nAll the details &amp;amp; book here https://t.co/rrZLdl
j1Vf https://t.co/8UhtrgoRsK', id_B=1513826526017564673, full_text_B='The Kent and Medway Business Summit returns
on Wednesday 27 April 2022 in the stunning Sibson building on the Canterbury campus of the University of Kent\n\n
All the details &amp;amp; book here https://t.co/Fl9sGU3L2c https://t.co/plApIlH9j0')]
```

In [235…

```
ls = [1512489448675876865, 1528829779100635136, 1518064686620721155]
news = dup_user.filter(dup_user.id_A.isin(ls)).limit(10).toPandas()
pd.set_option('display.max_rows', 30)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('max_colwidth', -1)
news[['user_name','full_text_A']]
```

```
/tmp/ipykernel_17136/2923367369.py:5: FutureWarning: Passing a negative integer is deprecated in version 1.0 and
will not be supported in future version. Instead, use None to not limit the column width.
  pd.set_option('max_colwidth', -1)
```

Out[235…

| | user_name | full_text_A |
|---|---|---|
| 0 | FoxNews | University professors unhappy by Florida Gov. DeSantis' attempt to hold faculty 'accountable' https://t.co/lHAUaoGaWO |
| 1 | ChannelRadio1 | The Kent and Medway Business Summit returns on Wednesday 27 April 2022 in the stunning Sibson building on the Canterbury campus of the University of Kent\n\nAll the details & book here https://t.co/rrZLdlj1Vf https://t.co/8UhtrgoRsK |
| 2 | BerkleyBearNews | Florida university must reinstate professor who was fired over 'Black privilege' tweets https://t.co/wvrzMdMrFU #news #topstories #berkleybearnews |

| | | |
|---|---|---|
| 3 | BerkleyBearNews | Florida university must reinstate professor who was fired over 'Black privilege' tweets https://t.co/wvrzMdMrFU #news #topstories #berkleybearnews |

## Social Media Influencers

In [194…
```
dup_user.filter(col('id_A')==1541928185960570880).limit(1).collect()
```

Out[194…
```
[Row(user_name='IndieBookButler', user_descrip='The best indie reads.\nAuthor pages, book launches, interviews, book boosts, and more! \nCome and find your next great read!', user_id=2981782721, id_A=1541928185960570880, full_text_A="Not Now, Katrin' picks up where 'The University Club - a Campus Affair' left off. This book includes eight original recipes, as well as passages intended for an adult (18+) audience https://t.co/KxIcTcPgR5 #fiction #humor #romance @artinchocolate", id_B=1589263878814318592, full_text_B="Not Now, Katrin' picks up where 'The University Club - a Campus Affair' left off. This book includes eight original recipes, as well as passages intended for an adult (18+) audience https://t.co/KxIcTcPgR5 #fiction #humor #romance @artinchocolate")]
```

In [196…
```
dup_user.filter(col('id_A')==1547761185323630592).limit(1).collect()
```

Out[196…
```
[Row(user_name='nohototasehena1', user_descrip='Web3 designer \u200d who is at the junction of web 3 acceptance and understanding, what a great time we live in ❤  #NFT #CryptoArt #NFTCommunity', user_id=1547760546212347904, id_A=1547761185323630592, full_text_A='My mum bought me a book on procrastination at the start of high school', id_B=1548150850027405312, full_text_B='My mum bought me a book on procrastination at the start of high school')]
```

In [208…
```
dup_user.filter(col('id_A')==1529543925081509888).limit(1).collect()
```

Out[208…
```
[Row(user_name='daypeacecomedy', user_descrip='For Booking: comediandayday@gmail.com  Podcast: Peace Talk ☮
“Peace of Mind (Side B)” comedy album debuted #1 on Amazon and iTunes!!  | □', user_id=764921841802579968, id_A=1529543925081509888, full_text_A='Florida State University. \nMiami Carol City High School. \nRogers State University. \nRosemary Anderson High School. \nWisconsin Lutheran High School. \nFrederick High School. \nTenaya Middle School. \nBethune-Cookman University. \nPershing Elementary School. \nWayne Community College.', id_B=1529611244604620800, full_text_B='Florida State University. \n\nMiami Carol City High School. \nRogers State University. \nRosemary Anderson High School. \nWisconsin Lutheran High School. \nFrederick High School. \nTenaya Middle School. \n
```

In [236…
```
ls = [1529543925081509888, 1547761185323630592, 1541928185960570880]
social = dup_user.filter(dup_user.id_A.isin(ls)).limit(10).toPandas()
pd.set_option('display.max_rows', 30)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('max_colwidth', -1)
social[['user_name','full_text_A']]
```

```
/tmp/ipykernel_17136/4042040084.py:5: FutureWarning: Passing a negative integer is deprecated in version 1.0 and will not be supported in future version. Instead, use None to not limit the column width.
  pd.set_option('max_colwidth', -1)
```

Out[236…

| | user_name | full_text_A |
|---|---|---|
| 0 | IndieBookButler | Not Now, Katrin' picks up where 'The University Club - a Campus Affair' left off. This book includes eight original recipes, as well as passages intended for an adult (18+) audience https://t.co/KxIcTcPgR5 #fiction #humor #romance @artinchocolate |
| 1 | IndieBookButler | Not Now, Katrin' picks up where 'The University Club - a Campus Affair' left off. This book includes eight original recipes, as well as passages intended for an adult (18+) audience https://t.co/KxIcTcPgR5 #fiction #humor #romance @artinchocolate |
| 2 | nohototasehena1 | My mum bought me a book on procrastination at the start of high school |
| 3 | nohototasehena1 | My mum bought me a book on procrastination at the start of high school |
| 4 | daypeacecomedy | Florida State University. \nMiami Carol City High School. \nRogers State University. \nRosemary Anderson High School. \nWisconsin Lutheran High School. \nFrederick High School. \nTenaya Middle School. \nBethune-Cookman University. \nPershing Elementary School. \nWayne Community College. |

## Educators

In [186…
```
dup_user.filter(col('id_A')==1551060261297086464).limit(1).collect()
```

Out[186…
```
[Row(user_name='SGaAngel67', user_descrip='Mom of 2 beautiful young ladies (1 by birth, 1 by ❤)\nGigi to 3 adorab
le boys\n□□\u200d□SpEd teacher in math\n□avid reader of MC/shifter/gay romance', user_id=705809602986778624, id_A=
551060261297086464, full_text_A='@kkirtley26 Any help would be greatly appreciated, trying to restock my 6th-12th
grade Special Education classroom; I primarily teach math in an alternative school setting. #clearthelist #cleart
helist2022 #AdoptATeacher #teachertwitter \n\nhttps://t.co/QeJheAl4Wo', id_B=1551069371598049280, full_text_B='@L
ynzforCongress Thank you for yor support, any help would be greatly appreciated. Trying to restock my 6th-12th gr
ade Special Education classroom; I primarily teach math in an alternative school setting. #clearthelist #clearthe
list2022 #AdoptATeacher #teachertwitter\n\nhttps://t.co/QeJheAl4Wo')]
```

In [184…
```
dup_user.filter(col('id_A')==1553133520935075840).limit(1).collect()
```

Out[184…
```
[Row(user_name='saragrieb', user_descrip='high school math teacher, former social worker. love kids & have a spec
ial ❤  for those who struggle Amazon wishlist https://tinyurl.com/GriebWishlist', user_id=25929137, id_A=15531335
20935075840, full_text_A='#clearthelist #mathteacher #clearthelist2022 high school math teacher in Missouri https
://t.co/DdA0ZGUiVA', id_B=1556099110712885248, full_text_B='@CapitalSatoshi High school math teacher')]
```

In [185…
```
dup_user.filter(col('id_A')==1578690237814820865).limit(1).collect()
```

Out[185…
```
[Row(user_name='CastlesofImagin', user_descrip='He/His/Him\nAuthor, Teacher, Randomness is fun.\nCastles of Imagi
nation series\nSuper Super Hero Store duology\nNational City Stories zines\n#InteractiveFiction', user_id=9258124
75433504769, id_A=1578690237814820865, full_text_A="@moordereht @mrsmoordereht My first published book was about
teenagers using RPGs to survive high school. It's #InteractiveFiction so you are the high schooler and their barb
arian character making choices about the frame tale #Gamebook.\nhttps://t.co/TtlqAim0zA", id_B=157869317828206182
4, full_text_B="@NMtDR_Magazine My first published book was about teenagers using RPGs to survive high school. It
's #InteractiveFiction so you are the high schooler and their barbarian character making choices about the frame
tale #Gamebook.\nhttps://t.co/TtlqAim0zA")]
```

In [237…
```
ls = [1578690237814820865, 1553133520935075840, 1551060261297086464]
educator = dup_user.filter(dup_user.id_A.isin(ls)).limit(10).toPandas()
pd.set_option('display.max_rows', 30)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('max_colwidth', -1)
educator[['user_name','full_text_A']]
```

```
/tmp/ipykernel_17136/419945617.py:5: FutureWarning: Passing a negative integer is deprecated in version 1.0 and w
ill not be supported in future version. Instead, use None to not limit the column width.
  pd.set_option('max_colwidth', -1)
```

Out[237…

| | user_name | full_text_A |
|---|---|---|
| 0 | saragrieb | #clearthelist #mathteacher #clearthelist2022 high school math teacher in Missouri https://t.co/DdA0ZGUiVA |
| 1 | CastlesofImagin | @moordereht @mrsmoordereht My first published book was about teenagers using RPGs to survive high school. It's #InteractiveFiction so you are the high schooler and their barbarian character making choices about the frame tale #Gamebook.\nhttps://t.co/TtlqAim0zA |
| 2 | SGaAngel67 | @kkirtley26 Any help would be greatly appreciated, trying to restock my 6th-12th grade Special Education classroom; I primarily teach math in an alternative school setting. #clearthelist #clearthelist2022 #AdoptATeacher #teachertwitter \n\nhttps://t.co/QeJheAl4Wo |
| 3 | SGaAngel67 | @kkirtley26 Any help would be greatly appreciated, trying to restock my 6th-12th grade Special Education classroom; I primarily teach math in an alternative school setting. #clearthelist #clearthelist2022 #AdoptATeacher #teachertwitter \n\nhttps://t.co/QeJheAl4Wo |

## Sports Platform

In [225…
```
dup_user.filter(col('id_A')==1575733638234222592).limit(1).collect()
```

Out[225…

```
[Row(user_name='cappertek', user_descrip='Make Smarter Sports Betting Picks and Predictions. Handicapping Tools,
Reviews, Betting Tips, and More! Free Sports Betting Pick Simulator ➡  https://t.co/ENvH4ZgNAK', user_id=16893798
37, id_A=1575733638234222592, full_text_A='NCAAF College Football Game Simulator: 10/1/2022 12:00 PM ET - Wake Fo
rest vs. Florida State Game Simulation and Free Picks Generated by Advanced Algorithms https://t.co/gOHUtjurrH',
id_B=1581301531436982272, full_text_B='NCAAF College Football Game Simulator: 7:30 PM ET - Clemson vs. Florida St
ate Game Simulation and Free Picks Generated by Advanced Algorithms https://t.co/nXEL2yAOaf')]
```

In [238…

```
ls = [1575733638234222592]
news = dup_user.filter(dup_user.id_A.isin(ls)).limit(10).toPandas()
pd.set_option('display.max_rows', 30)
pd.set_option('display.expand_frame_repr', False)
pd.set_option('max_colwidth', -1)
news[['user_name','full_text_A']]
```

```
/tmp/ipykernel_17136/1430154354.py:5: FutureWarning: Passing a negative integer is deprecated in version 1.0 and
will not be supported in future version. Instead, use None to not limit the column width.
  pd.set_option('max_colwidth', -1)
```

Out[238…

| | user_name | full_text_A |
|---|---|---|
| 0 | cappertek | NCAAF College Football Game Simulator: 10/1/2022 12:00 PM ET - Wake Forest vs. Florida State Game Simulation and Free Picks Generated by Advanced Algorithms https://t.co/gOHUtjurrH |

In [ ]: