Jeremey_test

July 30, 2021

1 Info

Original data: https://www.kaggle.com/reddit/reddit-comments-may-2015

We took first 10M of 54M Rows

[1]:
```python
import pandas as pd

from pyspark.sql import SparkSession
spark = SparkSession.builder.getOrCreate()

full_path = '/project/ds5559/r-slash-group8/sample.csv'

df = spark.read.csv(full_path,  inferSchema=True, header = True)
```

[2]:
```python
df.show(5)
```

```
+--------------------------------+-----------+----+------------+---------+------
----+------------+----------------------+-----------------+---------+-------+---
-----------+------+-----+--------+------------+-----+------------+--------------
------+-------------+------+----------------+----------+
|                             _c0|created_utc| ups|subreddit_id|  link_id|
name|score_hidden|author_flair_css_class|author_flair_text|subreddit|
id|removal_reason|gilded|downs|archived|      author|score|retrieved_on|
body|distinguished|edited|controversiality| parent_id|
+--------------------------------+-----------+----+------------+---------+------
----+------------+----------------------+-----------------+---------+-------+---
-----------+------+-----+--------+------------+-----+------------+--------------
------+-------------+------+----------------+----------+
|                               1| 1430438400|   4|
t5_378oi|t3_34di91|t1_cqug90g|           0|                    NA|
NA|soccer_jp|cqug90g|            NA|     0|    0|       0|       rx109|    4|
1432703079|                 |         null|  null|            null|      null|
|        |       null|null|        null|     null|      null|
null|                  null|             null|     null|   null|          null|
null| null|    null|        null| null|        null|                null|
null|  null|            null|      null|
|                  "|         NA|   0|           0|t3_34di91|     null|
null|                  null|             null|     null|   null|          null|
null| null|    null|        null| null|        null|                null|
null|  null|            null|      null|
|                               2| 1430438400|   4|
t5_2qo4s|t3_34g8mx|t1_cqug90h|          0|                  Heat|
Heat|      nba|cqug90h|            NA|     0|    0|       0|   WyaOfWade|    4|
1432703079|gg this one's ove…|           NA|     0|               0|
t3_34g8mx|
|                               3| 1430438400|   0|
t5_2cneq|t3_34f7mc|t1_cqug90i|          0|                    NA|
NA| politics|cqug90i|            NA|     0|    0|       0|Wicked_Truth|     0|
1432703079|Are you really im…|           NA|     0|
0|t1_cqufim0|
+--------------------------------+-----------+----+------------+----------+------
----+------------+----------------------+------------------+---------+-------+---
-----------+------+-----+--------+------------+-----+-------------+--------------
------+-------------+------+----------------+-----------+
only showing top 5 rows
```

[ ]: