Jordan_subset

July 30, 2021

0.1 Set Up

[2]:

```
import pandas as pd

from pyspark.ml import Pipeline
from pyspark.ml.classification import LogisticRegression
from pyspark.ml.feature import *

from pyspark.sql import SparkSession
from pyspark.sql.functions import col, countDistinct, udf
from pyspark.sql.types import ArrayType, IntegerType,  StringType

spark = SparkSession.builder.getOrCreate()
```

0.2 Pull in Data

[3]:

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

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

[4]:

```
# convert integer cols (ups, downs, and gilded) to integers
# Note: we could have done this by defining a schema before the csv read
df=df.withColumn("ups",df.ups.cast(IntegerType()))
df=df.withColumn("downs",df.downs.cast(IntegerType()))
df=df.withColumn("gilded",df.gilded.cast(IntegerType()))

# Confirm new schema
df.printSchema()
df.show(5)
```

```
root
 |-- _c0: string (nullable = true)
 |-- created_utc: string (nullable = true)
 |-- ups: integer (nullable = true)
 |-- subreddit_id: string (nullable = true)
 |-- link_id: string (nullable = true)
 |-- name: string (nullable = true)
 |-- score_hidden: string (nullable = true)
 |-- author_flair_css_class: string (nullable = true)
 |-- author_flair_text: string (nullable = true)
 |-- subreddit: string (nullable = true)
 |-- id: string (nullable = true)
 |-- removal_reason: string (nullable = true)
 |-- gilded: integer (nullable = true)
 |-- downs: integer (nullable = true)
 |-- archived: string (nullable = true)
 |-- author: string (nullable = true)
 |-- score: string (nullable = true)
 |-- retrieved_on: string (nullable = true)
 |-- body: string (nullable = true)
 |-- distinguished: string (nullable = true)
 |-- edited: string (nullable = true)
 |-- controversiality: string (nullable = true)
 |-- parent_id: string (nullable = true)

+--------------------------------+-----------+----+------------+---------+------
----+------------+----------------------+-----------------+---------+-------+---
-----------+------+-----+--------+------------+-----+------------+--------------
------+-------------+------+----------------+----------+
|                             _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
```

[5]:
```
df.count()
```

[5]:
```
15317725
```

[6]:
```
# Remove null in important columns
print(df.filter(df['body'].isNull()).count())
print(df.filter(df['subreddit_id'].isNull()).count())

df=df.filter(df['body'].isNotNull())
df=df.filter(df['subreddit_id'].isNotNull())

df.count()
```

```
5315303
2969558
```

[6]:
```
10002410
```

[7]:
```
# some rows have the body in the subreddit column, I'll remove these
print(df.filter(df['subreddit'].rlike('\s')).count())

df = df.filter(df['subreddit'].rlike('^[A-Za-z1-9_]+$'))
df.count() #filter out spaces and special characters except for underscores
```

```
5491
```

[7]:
```
9950519
```

[8]:
```
# how many subredits are there?
df.select('subreddit_id').distinct().count()
```

[8]:
```
26882
```

[9]:
```
# get highest volume subreddits
top_sr = df.groupby('subreddit_id').agg({'subreddit_id':'count'}).
 ↪sort(col('count(subreddit_id)').desc())
top_sr.show(5)
```

```
+------------+-------------------+
|subreddit_id|count(subreddit_id)|
+------------+-------------------+
|    t5_2qh1i|             756074|
|    t5_2rfxx|             194454|
|    t5_2qh0u|             142840|
|    t5_2qmg3|             138655|
|    t5_2qh33|             137526|
+------------+-------------------+
only showing top 5 rows
```

[10]:

```
# 1000 seems like a good cut off point.  Subreddits have 1000+ comments
top_sr.limit(1000).tail(1)
```

[10]:

```
[Row(subreddit_id='t5_2s4mv', count(subreddit_id)=1303)]
```

[11]:

```
# add names based on frist appearance
# top_sr = top_sr.withColumn('subreddit',df.
 ↪filter(df['subreddit_id']==top_sr['subreddit_id']))
top_sr = top_sr.join(df.select('subreddit','subreddit_id').
 ↪dropDuplicates(['subreddit']),on='subreddit_id', how='inner').
 ↪sort(col('count(subreddit_id)').desc())
top_sr


# NOTE: It looks like some of the data still has issues because I still have␣
 ↪duplicates probably meaning there are some rows that have comment text in␣
 ↪the subreddit column.
#They weren't eliminated because the comments happen to follow subreddit naming␣
 ↪conventions.  More work would be helpful here.
# My goal is to elminate unpopular subreddits so we have fewer categories to␣
 ↪work with.  Not sure how important working with names really is.
```

[11]:

```
DataFrame[subreddit_id: string, count(subreddit_id): bigint, subreddit: string]
```

[12]:

```
# top 1000 subreddits
top_sr2 = top_sr.limit(1000)
top_sr2.show(5)
```

```
+------------+-------------------+---------------+
|subreddit_id|count(subreddit_id)|      subreddit|
+------------+-------------------+---------------+
|    t5_2qh1i|             756074|      AskReddit|
|    t5_2rfxx|             194454|leagueoflegends|
|    t5_2qh0u|             142840|           pics|
|    t5_2qmg3|             138655|            nfl|
|    t5_2qh33|             137526|           funny|
+------------+-------------------+----------------+
only showing top 5 rows
```

[23]:
```python
# eliminate data from lower level records

df_topSr = df.join(top_sr2.select('subreddit_id'), on = 'subreddit_id')
```

[24]:
```python
df_topSr.columns
```

[24]:
```
['subreddit_id',
 '_c0',
 'created_utc',
 'ups',
 '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']
```

[25]:
```python
df_topSr.count()
```

[25]:
```
8799960
```

[26]:
```python
df_topSr.show(10)
```

```
+------------+---+-----------+---+---------+----------+------------+------------
----------+-----------------+---------------+-------+--------------+------+-----
+--------+---------------+-----+------------+--------------------+-------------+
------+----------------+----------+
|subreddit_id|_c0|created_utc|ups|  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|
+------------+---+-----------+---+---------+----------+------------+------------
----------+-----------------+---------------+-------+--------------+------+-----
+---------+--------------+-----+------------+--------------------+-------------+
--------+--------------+-----------+
|    t5_2qo4s|  2| 1430438400|  4|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|
|    t5_2cneq|  3| 1430438400|  0|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|
|    t5_2qh1i|  4| 1430438400|  3|t3_34f9rh|t1_cqug90j|           0|
NA|               NA|      AskReddit|cqug90j|             NA|     0|    0|
0|       jesse9o3|    3|  1432703079|No one has a Euro…|           NA|     0|
0|t1_cqug2sr|
|    t5_2qh1i|  5| 1430438400|  3|t3_34fvry|t1_cqug90k|           0|
NA|               NA|      AskReddit|cqug90k|             NA|     0|    0|
0| beltfedshooter|    3|  1432703079|"That the kid ""…|           NA|     0|
0| t3_34fvry|
|    t5_31k9i|  6| 1430438400|  1|t3_34gitq|t1_cqug90l|           0|
NA|               NA|     bloodborne|cqug90l|             NA|     0|    0|
0|     Rubenticus|    1|  1432703079|Haha, i was getti…|           NA|     0|
0|t1_cqug10q|
|    t5_2qjvn|  7| 1430438400|  6|t3_34fpen|t1_cqug90m|           0|
NA|               NA|  relationships|cqug90m|             NA|     0|    0|
0|silverraven1189|    6|  1432703079|After reading thi…|         null|  null|
null|      null|
|    t5_2s5fm|  8| 1430438400|  2|t3_34e7uq|t1_cqug90n|           0|
Titan3|               NA|Tennesseetitans|cqug90n|             NA|     0|    0|
0|     Scrubtanic|    2|  1432703079|Let's do this. Se…|    moderator|     0|
0| t3_34e7uq|
|    t5_2r090|  9| 1430438400|  6|t3_34gcwh|t1_cqug90o|           0|
T10B10|         [Philly]|         cigars|cqug90o|             NA|     0|    0|
0|       burnmyiz|    6|  1432703079|You can buy a mys…|           NA|     0|
0| t3_34gcwh|
|    t5_2sqho| 10| 1430438400|  5|t3_34gmag|t1_cqug90p|           0|
fan vp|   Virtus.pro Fan|GlobalOffensive|cqug90p|             NA|     0|    0|
0|      BEE_REAL_|    5|  1432703079|Nihilum and LG ar…|           NA|     0|
0| t3_34gmag|
|    t5_2qi5w| 11| 1430438400|  4|t3_34gmq6|t1_cqug90q|           0|
modernbird|             null|         eagles|cqug90q|             NA|     0|
0|       0|           SNVG|    4|  1432703079|      Fuck that what|
NA|     0|              0| t3_34gmq6|
+------------+---+-----------+---+---------+----------+------------+------------
----------+-----------------+---------------+-------+--------------+------+-----
+---------+--------------+-----+------------+--------------------+-------------+
------+----------------+----------+
only showing top 10 rows
```

[29]:

```
df_topSr.write.csv('jordan_subset', header=True)
```

[ ]:

```
```