ben

July 30, 2021

[1]:
```python
path = '/project/ds5559/r-slash-group8/sample.csv'
```

[2]:
```python
import pandas as pd
from pyspark.sql import SparkSession
import numpy as np
import re
import nltk
from pyspark.sql.functions import *
from pyspark.sql.types import ArrayType, IntegerType,  StringType, TimestampType
```

[3]:
```python
spark = SparkSession.builder.getOrCreate()
```

[4]:
```python
df_full = spark.read.csv(path,  inferSchema=True, header = True)
```

[5]:
```python
df_full.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
```

[6]:
```
df = df_full.
 ↪drop('_c0','subreddit_id','link_id','name','score_hidden','author_flair_css_class',␣
 ↪\
        'author_flair_text','archived','retrieved_on',␣
 ↪'edited','controversiality','parent_id','ups', \
       'downs', 'removal_reason', 'distinguished')
# keep score, subreddit, id, created_utc, author, body

df=df.withColumn("score",df.score.cast(IntegerType()))
df=df.withColumn("gilded",df.gilded.cast(IntegerType()))
df=df.withColumn("created_utc",from_unixtime(df.created_utc))
df=df.withColumnRenamed("id", "comment_id")
```

[7]:
```
df.show(5)
df.printSchema()
```

```
+-------------------+---------+----------+------+------------+-----+------------
--------+
|        created_utc|subreddit|comment_id|gilded|      author|score|
body|
+-------------------+---------+----------+------+------------+-----+------------
--------+
|2015-05-01 00:00:00|soccer_jp|   cqug90g|     0|       rx109|    4|
 |
|               null|     null|      null|  null|        null| null|
null|
|               null|     null|      null|  null|        null| null|
null|
|2015-05-01 00:00:00|      nba|   cqug90h|     0|   WyaOfWade|    4|gg this
```

```
one's ove…|
|2015-05-01 00:00:00| politics|   cqug90i|     0|Wicked_Truth|    0|Are you
really im…|
+-------------------+---------+----------+------+------------+-----+------------
--------+
only showing top 5 rows

root
 |-- created_utc: string (nullable = true)
 |-- subreddit: string (nullable = true)
 |-- comment_id: string (nullable = true)
 |-- gilded: integer (nullable = true)
 |-- author: string (nullable = true)
 |-- score: integer (nullable = true)
 |-- body: string (nullable = true)
```

[8]:

```python
df=df.filter(df['body'].isNotNull())
df=df.filter(df['subreddit'].isNotNull())
df=df.filter(df['comment_id'].isNotNull())
```

[9]:

```python
df.show()
```

```
+-------------------+------------------+----------+------+---------------+-----+
--------------------+
|        created_utc|         subreddit|comment_id|gilded|         author|score|
body|
+-------------------+------------------+----------+------+---------------+-----+
--------------------+
|2015-05-01 00:00:00|         soccer_jp|   cqug90g|     0|          rx109|    4|
 |
|2015-05-01 00:00:00|               nba|   cqug90h|     0|      WyaOfWade|
4|gg this one's ove…|
|2015-05-01 00:00:00|          politics|   cqug90i|     0|   Wicked_Truth|
0|Are you really im…|
|2015-05-01 00:00:00|         AskReddit|   cqug90j|     0|       jesse9o3|
3|No one has a Euro…|
|2015-05-01 00:00:00|         AskReddit|   cqug90k|     0| beltfedshooter|
3|"That the kid ""…|
|2015-05-01 00:00:00|        bloodborne|   cqug90l|     0|     Rubenticus|
1|Haha, i was getti…|
|2015-05-01 00:00:00|     relationships|   cqug90m|     0|silverraven1189|
6|After reading thi…|
|2015-05-01 00:00:00|   Tennesseetitans|   cqug90n|     0|     Scrubtanic|
2|Let's do this. Se…|
|2015-05-01 00:00:00|            cigars|   cqug90o|     0|       burnmyiz|
6|You can buy a mys…|
|2015-05-01 00:00:00|   GlobalOffensive|   cqug90p|     0|      BEE_REAL_|
5|Nihilum and LG ar…|
|2015-05-01 00:00:00|            eagles|   cqug90q|     0|           SNVG|    4|
Fuck that what|
|2015-05-01 00:00:00|         smashbros|   cqug90r|     0|       BiigLord|
1|Don't diss the Gr…|
|2015-05-01 00:00:00|      makinghiphop|   cqug90s|     0|        KingEze|
1|Your 16 bars seem…|
|2015-05-01 00:00:00|   GoogleCardboard|   cqug90t|     0|        cyborek|
3|Trinus vr is amaz…|
|2015-05-01 00:00:00|        offmychest|   cqug90u|     0|     Zekkystyle|
14|It's not your fau…|
|2015-05-01 00:00:00|               nrl|   cqug90v|     0|       jeauxoxo|
3|http://www.reddit…|
|2015-05-01 00:00:00|elderscrollsonline|   cqug90w|     0| Chainsaw_Ninja|
0|and what about al…|
|2015-05-01 00:00:00|         childfree|   cqug90x|     0|    tparkelaine|
5|I would be tempte…|
|2015-05-01 00:00:00|       ExNoContact|   cqug90y|     0|         vvvvfl|
1|I can't answer be…|
|2015-05-01 00:00:00|         AskReddit|   cqug90z|     0|  InterimFatGuy|    5|
NSFL|
+-------------------+------------------+----------+------+---------------+-----+
--------------------+
only showing top 20 rows
```

[10]:
```
df.count()
```

[10]:
```
10002410
```

[11]:
```
df_reduced = df.filter((df['subreddit'] == 'AskReddit') | (df['subreddit'] ==␣
 ↪'nfl'))
df_reduced = df_reduced.withColumn('subreddit_bin',when(df['subreddit'] ==␣
 ↪'AskReddit',0).otherwise(1))
df_reduced.show()
```

```
+-------------------+---------+----------+------+--------------+-----+----------
----------+-------------+
|        created_utc|subreddit|comment_id|gilded|        author|score|
body|subreddit_bin|
+-------------------+---------+----------+------+--------------+-----+----------
----------+-------------+
|2015-05-01 00:00:00|AskReddit|   cqug90j|     0|      jesse9o3|    3|No one has
a Euro…|            0|
|2015-05-01 00:00:00|AskReddit|   cqug90k|     0|beltfedshooter|    3|"That the
kid ""…|            0|
|2015-05-01 00:00:00|AskReddit|   cqug90z|     0| InterimFatGuy|    5|
NSFL|            0|
|2015-05-01 00:00:01|AskReddit|   cqug91c|     0|   JuanTutrego|    1|I'm a guy
and I h…|            0|
|2015-05-01 00:00:01|AskReddit|   cqug91e|     0|   dcblackbelt|  101|Mid
twenties male…|            0|
|2015-05-01 00:00:02|AskReddit|   cqug920|     0| TheDoorsShirt|    1|Fran
Drescher lau…|            0|
|2015-05-01 00:00:02|AskReddit|   cqug921|     0|youthfulvictim|   -5|I honestly
wouldn…|            0|
|2015-05-01 00:00:02|AskReddit|   cqug923|     0|   sonovadoyle|    1|
&lt;/3 |            0|
|2015-05-01 00:00:02|AskReddit|   cqug929|     0|      boludo54|    1|no money,
just ID…|            0|
|2015-05-01 00:00:03|      nfl|   cqug92m|     0|     Stokest26|    4|"Do you
get to ju…|            1|
|2015-05-01 00:00:03|AskReddit|   cqug92o|     0| mister_sleepy|    1|Smoking
tobacco. …|            0|
|2015-05-01 00:00:03|      nfl|   cqug92p|     0|    Drakengard|
2|https://i.imgur.c…|            1|
|2015-05-01 00:00:03|AskReddit|   cqug92s|     0|baconinstitute|    1|   Bootleg
Fireworks|            0|
|2015-05-01 00:00:03|AskReddit|   cqug92v|     0|   cpmustang90|    1|I'm a
member of t…|            0|
|2015-05-01 00:00:03|AskReddit|   cqug937|     0|       hamhead|    1|That's one
reason…|            0|
|2015-05-01 00:00:03|      nfl|   cqug93g|     0|     wesman212|    1|Can we
sticky the…|            1|
|2015-05-01 00:00:03|AskReddit|   cqug93k|     0|  mrjosemeehan|    7|Go back to
Route …|            0|
|2015-05-01 00:00:04|AskReddit|   cqug93o|     0|  Peacehamster|    2|They're
oh-so-bra…|            0|
|2015-05-01 00:00:04|AskReddit|   cqug93u|     0|    bunnylumps|   20|"I think
the whol…|            0|
|2015-05-01 00:00:04|AskReddit|   cqug93v|     0|       DawkTux|    1|Fuck pound
coins…|            0|
+-------------------+---------+----------+------+--------------+-----+----------
----------+-------------+
only showing top 20 rows
```

[12]:
```
df_reduced.count()
```

[12]: 894729

[13]:
```
df_train, df_test = df_reduced.randomSplit([0.75,0.25])
```

[14]:
```
df_train.count()
```

[14]: 671189

[15]:
```python
df_train.groupBy('subreddit_bin').count().show()
```

```
+-------------+-------+
|subreddit_bin| count|
+-------------+-------+
|            1|103800|
|            0|567389|
+-------------+-------+
```

[ ]:
```python

```

[16]:
```python
df_test.groupBy('subreddit_bin').count().show()
```

```
+-------------+-------+
|subreddit_bin| count|
+-------------+-------+
|            1| 34855|
|            0|188685|
+-------------+-------+
```

[17]:
```python
from pyspark.ml import Pipeline
from pyspark.ml.feature import *
from pyspark.ml.classification import LogisticRegression
```

[18]:
```python
tok = Tokenizer(inputCol="body", outputCol="body_tokens")
rem = StopWordsRemover(inputCol="body_tokens", outputCol="tokens_filtered")
htf = HashingTF(numFeatures = 4096, inputCol="tokens_filtered", outputCol="tf")
w2v = Word2Vec(inputCol="body_tokens", outputCol="w2v")
va = VectorAssembler(inputCols=["tf", "w2v"], outputCol="features")
lr = LogisticRegression(labelCol='subreddit_bin', featuresCol='features',␣
 ↪maxIter=10, regParam=0.01)

pipeline = Pipeline(stages=[tok, rem, htf, w2v, va, lr])
```

[19]:
```python
model = pipeline.fit(df_train)
```

[21]:
```python
prediction = model.transform(df_test)
prediction.show()
```

```
+-------------------+---------+----------+------+--------------+-----+----------
----------+-------------+--------------------+--------------------+------------
-------+--------------------+--------------------+--------------------+---------
-----------+----------+
|        created_utc|subreddit|comment_id|gilded|        author|score|
body|subreddit_bin|         body_tokens|     tokens_filtered|
tf|                 w2v|            features|       rawPrediction|
probability|prediction|
+-------------------+---------+----------+------+--------------+-----+----------
----------+-------------+--------------------+--------------------+-------------
-------+--------------------+--------------------+--------------------+---------
-----------+----------+
|2015-05-01 00:00:00|AskReddit|   cqug90j|     0|      jesse9o3|    3|No one has
a Euro…|            0|[no, one, has, a,…|[one, european, a…|(4096,[99,1343
,17…|[0.11656655858016…|(4196,[99,1343,17…|[4.44445710690465…|[0.9883928
2882496…|       0.0|
|2015-05-01 00:00:02|AskReddit|   cqug920|     0| TheDoorsShirt|    1|Fran
Drescher lau…|            0|[fran, drescher, …|[fran, drescher, …|(4096,[2
089,2781,…|[0.01728342659771…|(4196,[2089,2781,…|[2.37812174586286…|[0.9
1514369078506…|       0.0|
|2015-05-01 00:00:03|AskReddit|   cqug92v|     0|   cpmustang90|    1|I'm a
member of t…|            0|[i'm, a, member, …|[member, tea, par…|(4096,[79
3,4029,4…|[0.09982642637831…|(4196,[793,4029,4…|[4.21165572837286…|[0.98
539467025667…|       0.0|
|2015-05-01 00:00:03|AskReddit|   cqug93k|     0|  mrjosemeehan|    7|Go back to
Route …|            0|[go, back, to, ro…|[go, back, route,…|(4096,[487,119
8,1…|[-0.1260713236406…|(4196,[487,1198,1…|[0.71144015847846…|[0.6707193
0405816…|       0.0|
|2015-05-01 00:00:03|      nfl|   cqug92m|     0|     Stokest26|    4|"Do you
get to ju…|            1|["do, you, get, t…|["do, get, make, …|(4096,[716,
1157,1…|[-0.0254123289890…|(4196,[716,1157,1…|[1.68832637905484…|[0.8440
0393562142…|       0.0|
|2015-05-01 00:00:04|AskReddit|   cqug93o|     0|  Peacehamster|    2|They're
oh-so-bra…|            0|[they're, oh-so-b…|[oh-so-bravely, s…|(4096,[1465
,2470,…|[-0.0683531030697…|(4196,[1465,2470,…|[3.95176178321757…|[0.9811
4166345578…|       0.0|
|2015-05-01 00:00:04|AskReddit|   cqug93u|     0|    bunnylumps|   20|"I think
the whol…|            0|["i, think, the, …|["i, think, whole…|(4096,[193,3
08,43…|[-0.0194499432691…|(4196,[193,308,43…|[2.11907616335345…|[0.89274
350189698…|       0.0|
|2015-05-01 00:00:04|AskReddit|   cqug93v|     0|       DawkTux|    1|Fuck pound
coins…|            0|[fuck, pound, coi…|[fuck, pound, coi…|(4096,[123,156
1,1…|[-0.0110021372410…|(4196,[123,1561,1…|[2.65875610691634…|[0.9345486
2186456…|       0.0|
|2015-05-01 00:00:04|AskReddit|   cqug94d|     0|    SpareLiver|   11|"Yes,
that's what…|            0|["yes,, that's, w…|["yes,, going, on…|(4096,[14
0,843,85…|[0.01932223575810…|(4196,[140,843,85…|[3.27330875899342…|[0.96
350170671326…|       0.0|
|2015-05-01 00:00:04|      nfl|   cqug93x|     0|      BirdLaw_|    2|Woo. 2015
NFL Dra…|            1|[woo., 2015, nfl,…|[woo., 2015, nfl,…|(4096,[679,73
4,81…|[-0.0013861323241…|(4196,[679,734,81…|[-9.2637970782738…|[9.478571
28419531…|       1.0|
|2015-05-01 00:00:07|      nfl|   cqug95w|     0|     kmhines88|    2|In a show
```

```
of frie…|            1|[in, a, show, of,…|[show, friendship…|(4096,[1522,1
578,…|[0.01060863945167…|(4196,[1522,1578,…|[-6.0436342928132…|[0.002367
30192353…|       1.0|
|2015-05-01 00:00:07|      nfl|   cqug960|     0|   Semper-Fido|    1|Whatchu
doing her…|            1|[whatchu, doing, …| [whatchu, hawkeye?]|(4096,[3100
,3807]…|[0.00932125747203…|(4196,[3100,3807,…|[1.79152247754273…|[0.8571
1383529654…|       0.0|
|2015-05-01 00:00:08|      nfl|   cqug961|     0|     bwburke94|   12|
#FUCK GOODELL|            1|    [#fuck, goodell]|    [#fuck, goodell]|(4096,[154
0,3186]…|[-0.0096676694229…|(4196,[1540,3186,…|[-3.3474090357063…|[0.033
98011114624…|       1.0|
|2015-05-01 00:00:09|AskReddit|   cqug981|     0|  haggardclint|    1|When I was
a youn…|            0|[when, i, was, a,…|[youngster, teens…|(4096,[192,314
,34…|[-0.0179296344703…|(4196,[192,314,34…|[4.28133498207571…|[0.9863643
0775434…|       0.0|
|2015-05-01 00:00:09|      nfl|   cqug97x|     0|    TheAquaman|    2|If we fuck
this u…|            1|[if, we, fuck, th…|       [fuck, up…]|(4096,[166,363
9],…|[-0.1451669842004…|(4196,[166,3639,4…|[0.51484269706837…|[0.6259410
2776761…|       0.0|
|2015-05-01 00:00:10|AskReddit|   cqug98u|     0|   Gum_Disease|    1|I don't
know how …|            0|[i, don't, know, …|[know, read, comm…|(4096,[169,
282,32…|[-0.0456981721384…|(4196,[169,282,32…|[8.10439154574620…|[0.9996
9788224795…|       0.0|
|2015-05-01 00:00:12|AskReddit|   cqug9an|     0|  7LeagueBoots|    1|Here is
how all […|            0|[here, is, how, a…|[[those, idiots](…|(4096,[338,
2828,3…|[-0.0570049323141…|(4196,[338,2828,3…|[1.53642857728156…|[0.8229
4494554689…|       0.0|
|2015-05-01 00:00:12|      nfl|   cqug9a7|     0| smallgiantman|    2|Tampa Bay
has gon…|            1|[tampa, bay, has,…|[tampa, bay, gone…|(4096,[205,41
9,12…|[-0.0797286167691…|(4196,[205,419,12…|[-1.8013828562423…|[0.141682
81407853…|       1.0|
|2015-05-01 00:00:13|      nfl|   cqug9bh|     0| zombiebillnye|    3|I want
what you'r…|            1|[i, want, what, y…|    [want, smoking.]|(4096,[184
0,2264]…|[-0.0141300449147…|(4196,[1840,2264,…|[2.76124174528225…|[0.940
54511042451…|       0.0|
|2015-05-01 00:00:14|AskReddit|   cqug9bs|     0|JimmySmackCorn|    2|10000
DVD's of Pa…|            0|[10000, dvd's, of…|[10000, dvd's, pa…|(4096,[13
49,1957,…|[0.03559910033696…|(4196,[1349,1957,…|[4.27612981068178…|[0.98
629412192294…|       0.0|
+-------------------+---------+----------+------+--------------+-----+----------
----------+-------------+--------------------+--------------------+-------------
-------+--------------------+--------------------+--------------------+---------
-----------+----------+
only showing top 20 rows
```

[31]:

```
prediction.groupBy('prediction').count().show()
```

```
+----------+------+
|prediction| count|
+----------+------+
|       0.0|205459|
|       1.0| 18061|
+----------+------+
```

[39]:

```
predictionsAndLabels = prediction.drop('created_utc', 'subreddit',␣
 ↪'comment_id', 'gilded', 'author', 'score', 'body', 'body_tokens',␣
 ↪'tokens_filtered', 'rawPrediction', 'probability', 'tf', 'w2v', 'features')
```

[45]:

```
predictionsAndLabels = predictionsAndLabels.withColumnRenamed('subreddit_bin',␣
 ↪'label')
predictionsAndLabels = predictionsAndLabels.withColumn('label',␣
 ↪predictionsAndLabels.label.cast('float'))
predictionsAndLabels = predictionsAndLabels.withColumn('prediction',␣
 ↪predictionsAndLabels.prediction.cast('float'))
predictionsAndLabels.show()
```

```
+-----+----------+
|label|prediction|
+-----+----------+
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  1.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  1.0|       1.0|
|  1.0|       1.0|
|  1.0|       0.0|
|  1.0|       1.0|
|  0.0|       0.0|
|  1.0|       0.0|
|  0.0|       0.0|
|  0.0|       0.0|
|  1.0|       1.0|
|  1.0|       0.0|
|  0.0|       0.0|
+-----+----------+
only showing top 20 rows
```

[46]:

```
from pyspark.mllib.evaluation import MulticlassMetrics
```

[47]:

```
rdd = predictionsAndLabels.rdd.map(tuple)
```

[48]:

```
metrics = MulticlassMetrics(rdd)
```

[49]:

```
metrics.confusionMatrix().toArray()
```

[49]:

```
array([[185928.,  19345.],
       [  2757.,  15510.]])
```

Model biased toward predicting /r/Askreddit, could be tuned.

[ ]:

```

```