data-merge

```
%pyspark
part_file_dir = '/user/tw2770_nyu_edu/final-project/chunk-merged'
emotion = spark.read.parquet(part_file_dir)

emotion.count()
```

SPARK JOB FINISHED

```
20634
```

Took 2 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:17:58 PM.

```
%pyspark
emotion.show()
```

SPARK JOB (http://nyu-dataproc-sw-cfv5.c.hpc-dataproc-19b8.internal:46637/jobs/job?id=36) FINISHED

```
+-------+-----------+----------+----------+----------+-----------+-----------+
|     id|    sadness|       joy|      love|     anger|       fear|   surprise|
+-------+-----------+----------+----------+----------+-----------+-----------+
|2332484|  0.8546097|0.009181495|0.008009075| 0.03183697|  0.09305706|0.0033056184|
|2421428|0.0028901747|0.003345478|0.002211651| 0.98318946|  0.00749467| 8.684348E-4|
|1427541| 0.049701095|  0.8126464|0.030297162| 0.09761354|0.0073109404|0.0024308232|
|3071117| 0.043006115|0.073656656|0.012211307|  0.7528574|  0.11225884| 0.006009794|
| 766257| 0.016775277| 0.02016584|0.008843849| 0.81806326|  0.13238029|0.0037714203|
|1227797|  0.48852038| 0.14427386|0.029510187|   0.312028|  0.02164123|0.0040263017|
| 855344| 0.012327597|0.005283105|0.002541368|  0.9598838| 0.019165717| 7.983267E-4|
| 883595| 0.100263976| 0.41116253|0.031876218| 0.40993154| 0.039036658|0.0077291164|
|2061180| 0.015190468| 0.14963141|0.010143468| 0.79161555| 0.029288875|0.0041302955|
|4197410|0.0052977474| 0.13567276|  0.4404063|  0.3897522| 0.022303866| 0.006567108|
|1105989| 0.058191113|  0.5782798| 0.06114255| 0.27934405| 0.019050822|0.0039917836|
|1639885|  0.23531379|0.010899221|0.004241868|  0.6900049| 0.056897115|0.0026431484|
|3074512|   0.1577825|  0.1719081| 0.50533485| 0.14523332|0.0151718585|0.0045693796|
|1501806|   0.8427653| 0.10454262|0.022209354|0.025028888|0.0036526953|0.0018011202|
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:18:01 PM.

Find the emotion with the highest score

FINISHED

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:21:41 PM.

```
%pyspark
from pyspark import StorageLevel
from pyspark.sql import Window, SparkSession
from pyspark.sql.functions import udf, col, regexp_replace, lower, greatest, ntile, when
from pyspark.sql.types import StringType, ArrayType, StructType, StructField, FloatType


emotion = emotion.withColumn("max_score", greatest("sadness", "joy", "love", "anger", "fea
                    .withColumn("emotion",
                                when(col("max_score") == col("sadness"), "sadness")
                                .when(col("max_score") == col("joy"), "joy")
                                .when(col("max_score") == col("love"), "love")
                                .when(col("max_score") == col("anger"), "anger")
                                .when(col("max_score") == col("fear"), "fear")
                                .when(col("max_score") == col("surprise"), "surprise")
                                .otherwise("unknown"))
```

FINISHED

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:18:05 PM.

```
%pyspark
emotion.show()
```

SPARK JOB (http://nyu-dataproc-sw-cfv5.c.hpc-dataproc-19b8.internal:46637/jobs/job?id=37) FINISHED

data-merge

```
+-------+-----------+----------+----------+----------+-----------+-----------+-------
---+-------+
|     id|    sadness|       joy|      love|     anger|       fear|   surprise| max_sc
ore|emotion|
+-------+-----------+----------+----------+----------+-----------+-----------+-------
---+-------+
|2332484|  0.8546097|0.009181495|0.008009075| 0.03183697|  0.09305706|0.0033056184| 0.8546
097|sadness|
|2421428|0.0028901747|0.003345478|0.002211651| 0.98318946|  0.00749467| 8.684348E-4|0.98318
946|  anger|
|1427541| 0.049701095|  0.8126464|0.030297162| 0.09761354|0.0073109404|0.0024308232| 0.8126
464|    joy|
|3071117| 0.043006115|0.073656656|0.012211307|  0.7528574|  0.11225884| 0.006009794| 0.7528
574|  anger|
| 766257| 0.016775277| 0.02016584|0.008843849| 0.81806326|  0.13238029|0.0037714203|0.81806
326|  anger|
|1227797|  0.48852038| 0.14427386|0.029510187|   0.312028|  0.02164123|0.0040263017|0.48852
```

Took 1 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:18:08 PM.

Drop unused columns

FINISHED

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:22:39 PM.

```
%pyspark

columns_to_drop = ["sadness", "joy", "love", "anger", "fear", "surprise", "max_score"]
emotion = emotion.drop(*columns_to_drop).persist()

emotion.show()
```

SPARK JOB FINISHED

```
+-------+-------+
|     id|emotion|
+-------+-------+
|2332484|sadness|
|2421428|  anger|
|1427541|    joy|
|3071117|  anger|
| 766257|  anger|
|1227797|sadness|
| 855344|  anger|
| 883595|    joy|
|2061180|  anger|
|4197410|   love|
|1105989|    joy|
|1639885|  anger|
|3074512|   love|
|1501806|sadness|
```

Took 2 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:18:19 PM.

Calculate the distribution among the emotions

FINISHED

data-merge

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:23:17 PM.

```
%pyspark

# Group by 'emotion' and count occurrences
emotion_counts = emotion.groupBy("emotion").count()
emotion_counts.show()
```

SPARK JOB FINISHED

```
+--------+-----+
| emotion|count|
+--------+-----+
|     joy| 5579|
|    love|  862|
|    fear|  812|
|   anger|10725|
| sadness| 2516|
|surprise|  140|
+--------+-----+
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:18:24 PM.

```
%pyspark

from pyspark.sql.functions import col, round

# Calculate total number of records
total = emotion.count()

# Add a 'percentage' column
emotion_distribution = emotion_counts.withColumn(
    "percentage",
    round((col("count") / total) * 100, 2)
).orderBy(col("count").desc())

# Show the distribution with percentages
emotion_distribution.show()
```

SPARK JOB FINISHED

```
+--------+-----+----------+
| emotion|count|percentage|
+--------+-----+----------+
|   anger|10725|     51.98|
|     joy| 5579|     27.04|
| sadness| 2516|     12.19|
|    love|  862|      4.18|
|    fear|  812|      3.94|
|surprise|  140|      0.68|
+--------+-----+----------+
```

Took 1 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:18:30 PM.

```
%pyspark

df_path = '/user/tw2770_nyu_edu/final-project/lyrics_partitioned.parquet'
```

SPARK JOB FINISHED

```
df = spark.read.parquet(df_path)
df.show()
```

```
+--------------------+-------+-----------------+----+-----+--------------------+----------
----------+----+--------+--------------+--------------------+
|               title|    tag|           artist|year|views|            features|
lyrics|  id|language|views_partition|      lyrics_cleaned|
+--------------------+-------+-----------------+----+-----+--------------------+----------
----------+----+--------+--------------+--------------------+
|       A poco a poco|    pop|       Mia Martini|2003|   28|                  {}|Ti ho fatt
o cresc...|NULL|    NULL|     medium-low|ti ho fatto cresc...|
|Ive Been Up These...|    pop|              Gob|2003|  353|                  {}|It was the
worst ...|NULL|    NULL|    medium-high|it was the worst ...|
| Im So Hot Interlude|    pop|     Paris Bennett|2007|   62|                  {}|
Hold up|NULL|    NULL|     medium-low|             hold up|
|Contigo En La Dis...|    pop|    Caetano Veloso|1994|  191|                  {}|No existe
un mome...|NULL|    NULL|    medium-high|no existe un mome...|
|           I Am Siam|    pop|        The Grates|2006|  123|                  {}|I am Siam,
oh, he...|NULL|    NULL|    medium-high|i am siam oh hear...|
|              X-Town|    pop|     Roger Chapman|2015|   33|                  {}|Itza binsa
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:18:52 PM.

Merge `emotion` data with original dataset by `id` column

FINISHED

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:24:30 PM.

```
%pyspark
from pyspark.sql.functions import broadcast

df_merged = emotion.join(broadcast(df), on="id", how="inner").persist()
df_merged.show()
```

SPARK JOB FINISHED

```
+-------+-------+--------------------+----+--------------------+----+-----+----------------
----+--------------------+--------+--------------+--------------------+
|     id|emotion|               title| tag|              artist|year|views|            feat
ures|              lyrics|language|views_partition|      lyrics_cleaned|
+-------+-------+--------------------+----+--------------------+----+-----+----------------
----+--------------------+--------+--------------+--------------------+
|2332484|sadness|             Sinners|rock|         LGND (Band)|2015|  524|
{}|The Dirty kind li...|      en|          high|the dirty kind li...|
|2421428|  anger|           Computers| rap|        Young DayDay|2016|  525|
{}|[Verse 1:Young Da...|      en|          high|verse 1young dayd...|
|1427541|    joy|Drinking Song Fro...| pop|    Rudimentary Peni|2015|  526|
{}|Come hither, my l...|      en|          high|come hither my la...|
|3071117|  anger|          Long Grind| rap|       Ryder (Rider)|2017|  526|
{}|(Chorus) Get on t...|      en|          high|chorus get on the...|
| 766257|  anger|Drunk Girl At The...| pop|Reggie and the Fu...|2003|  528|
{}|"So I start telli...|      en|          high|so i start tellin...|
|1227797|sadness|     Feel Their Pain| pop|              Insted|2015|  529|
```

Took 5 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:19:07 PM.

```
%pyspark

columns_to_drop = ["year", "features", "language", "lyrics_cleaned", "views_partition"]
final_df= df_merged.drop(*columns_to_drop).persist()
```

SPARK JOB FINISHED

```
final_df.show()
```

```
+-------+-------+--------------------+----+--------------------+-----+--------------------+
|     id|emotion|               title| tag|              artist|views|              lyrics|
+-------+-------+--------------------+----+--------------------+-----+--------------------+
|2332484|sadness|             Sinners|rock|         LGND (Band)|  524|The Dirty kind li...|
|2421428|  anger|           Computers| rap|        Young DayDay|  525|[Verse 1:Young Da...|
|1427541|    joy|Drinking Song Fro...| pop|    Rudimentary Peni|  526|Come hither, my l...|
|3071117|  anger|          Long Grind| rap|       Ryder (Rider)|  526|(Chorus) Get on t...|
| 766257|  anger|Drunk Girl At The...| pop|Reggie and the Fu...|  528|"So I start telli...|
|1227797|sadness|     Feel Their Pain| pop|              Insted|  529|Take your place a...|
| 855344|  anger|           Salt Mine| pop|               Assck|  530|There is prostitu...|
| 883595|    joy|     Song For Audrey| pop|    Backseat Goodbye|  530|If I'm Frank Sina...|
|2061180|  anger|             Kaptein| pop|         Kurt Darren|  532|Kaptein - span di...|
|4197410|   love|Escort Services i...| rap|         Anisa Goyal|  532|Ahmedabad service...|
|1105989|    joy|  Smile in Your Eyes| pop|          Tom Lehman|  533|Hey child, how yo...|
|1639885|  anger|    Put a Sock in It|rock|     Killwhitneydead|  533|Make sure you get...|
|3074512|   love|          Bewildered|rock|For The Likes Of You|  533|Lost in my mind n...|
|1501806|sadness|         Sing a Song| pop|          Statistics|  534|The songs are all...|
|1056033|  anger|        Ultra Twist|rock|          The Cramps|  534|Hey, let's twist ...|
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:19:13 PM.

Create `rating` column from 1 to 10 by the percentile of `views` column

FINISHED

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:25:29 PM.

```
%pyspark
from pyspark.sql import Window

window_spec = Window.orderBy("views")
final_df = final_df.withColumn("rating", ntile(10).over(window_spec))

final_df.show()
```

SPARK JOB FINISHED

```
+-------+-------+--------------------+----+--------------------+-----+--------------------+
------+
|     id|emotion|               title| tag|              artist|views|              lyrics|
rating|
+-------+-------+--------------------+----+--------------------+-----+--------------------+
------+
| 681265|    joy|Efficient Home Re...|misc|General Contracto...|    0|Want a great home...|
1|
| 689768|    joy|Honor Power and A...|misc|              Cjasso|    0|I wish I have the...|
1|
| 425813|   love|0expatriate0's "R...|misc|        0expatriate0|    0|The young compose...|
1|
| 461670|    joy|               PARTY|misc|      Emmanuel Ellis|    0|I like to party e...|
1|
| 486990|    joy|Draft CONSTRUCTED...|misc|      Samuel McAuley|    0|With the increasi...|
1|
| 528003|    joy|Creation and arch...|misc|            Andrear3|    0|The genesis and G...|
1|
```

Took 1 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:19:18 PM.

```
%pyspark
```

SPARK JOB FINISHED

```
final_df = final_df.drop("views").persist()

final_df.show()
```

data-merge

```
+-------+-------+--------------------+----+--------------------+--------------------+------
+
|     id|emotion|               title| tag|              artist|              lyrics|rating
|
+-------+-------+--------------------+----+--------------------+--------------------+------
+
| 681265|    joy|Efficient Home Re...|misc|General Contracto...|Want a great home...|     1
|
| 689768|    joy|Honor Power and A...|misc|              Cjasso|I wish I have the...|     1
|
| 425813|   love|0expatriate0’s “R...|misc|        0expatriate0|The young compose...|     1
|
| 461670|    joy|               PARTY|misc|      Emmanuel Ellis|I like to party e...|     1
|
| 486990|    joy|Draft CONSTRUCTED...|misc|      Samuel McAuley|With the increasi...|     1
|
| 528003|    joy|Creation and arch...|misc|            Andrear3|The genesis and G...|     1
+
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:19:22 PM.

```
%pyspark
final_df = final_df.dropDuplicates(["id", "title", "tag"])
print("Number of unique songs: ", final_df.count())
final_df.show(20)
```

SPARK JOB FINISHED

```
Number of unique songs:  20637
+-------+-------+--------------------+-------+--------------------+--------------------+---
---+
|     id|emotion|               title|    tag|              artist|              lyrics|rat
ing|
+-------+-------+--------------------+-------+--------------------+--------------------+---
---+
|    000|   love|The Endless Begin...|   misc|    Zecharia Sitchin|"Of the evidence ...|
5|
|    000|   love|       You cheated..|    rap|          Miley sims|"You did somethin...|
2|
|1000159|sadness|       Until The End|    pop|                 Six|We stood together...|
4|
|1000542|  anger|    Hasten the Blows|    pop|     Bluebottle Kiss|"Sleeping in to t...|
8|
|1000626|  anger|              Gitana|    pop|        Daniela Romo|Raices en el alma...|
8|
|1000633|sadness|    See You Tomorrow|    pop|             T.S.O.L|Animals and fasci...|
```

Took 1 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:36:25 PM.

Partition By Top Keywords

FINISHED

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:20:45 PM.

```
%pyspark
from pyspark.sql.functions import col, array, when, lit, explode, expr

# Step 1: Define the list of keywords
```

SPARK JOB FINISHED

```
keywords = [
    'like', 'life', 'love', 'yeah', 'shit', 'fuck', 'bitch', 'world', 'mind', 'heart', 'gi
        'night', 'every night', 'need loving', 'nothing left', 'yo bitch', 'say goodbye',
        hop', 'say love', 'leave alone', 'never stop', 'never forget', 'years ago', 'juicy
]
```

data-merge

```
# Step 2: Create a column containing an array of matched keywords
keywords_df = final_df.withColumn(
    "matched_keywords",
    array(*[when(col("lyrics").contains(keyword), lit(keyword)) for keyword in keywords])
)

# Step 3: Remove nulls from the `matched_keywords` array
keywords_df = keywords_df.withColumn(
    "matched_keywords",
    expr("filter(matched_keywords, x -> x IS NOT NULL)")
)

print("Count of rows before exploding by keywords", keywords_df.count())

# Step 4: Explode the `matched_keywords` column to create one row per keyword
keywords_df = keywords_df.withColumn("keyword", explode(col("matched_keywords")))

print("Count of rows after exploding by keywords", keywords_df.count())
print("Number of unique songs with lyrics containing top keywords", keywords_df.select("Ti
```

```
Count of rows before exploding by keywords 20637
Count of rows after exploding by keywords 45825
Number of unique songs with lyrics containing top keywords 14447
+-------+-------+--------------------+----+---------------+--------------------+------+---
-----------------+---------+
|     id|emotion|               title| tag|          artist|              lyrics|rating|
matched_keywords|  keyword|
+-------+-------+--------------------+----+---------------+--------------------+------+---
-----------------+---------+
|    000|   love|The Endless Begin...|misc|Zecharia Sitchin|"Of the evidence ...|     5|[li
ke, life, year...|     like|
|    000|   love|The Endless Begin...|misc|Zecharia Sitchin|"Of the evidence ...|     5|[li
ke, life, year...|     life|
|    000|   love|The Endless Begin...|misc|Zecharia Sitchin|"Of the evidence ...|     5|[li
ke, life, year...|years ago|
|    000|   love|       You cheated..| rap|      Miley sims|"You did somethin...|     2| [l
ike, love, night]|     like|
```

Took 5 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:36:53 PM.

%pyspark SPARK JOB FINISHED

```
# Drop duplicates and only select the needed columns
result_df = keywords_df.dropDuplicates(["id", "rating", "emotion", "keyword"])
result_df = result_df.select("title", "tag", "artist", "rating", "emotion", "keyword")
print("Count of rows after droping duplicates:", result_df.count())
print("Number of unique books with review containing top keywords", result_df.select("Titl
result_df.show(20)
```

```
Count of rows after droping duplicates: 45825
Number of unique books with review containing top keywords 14447
+--------------------+----+---------------+------+-------+---------+
|               title| tag|         artist|rating|emotion|  keyword|
+--------------------+----+---------------+------+-------+---------+
|       You cheated..| rap|      Miley sims|     2|   love|     like|
|       You cheated..| rap|      Miley sims|     2|   love|     love|
|       You cheated..| rap|      Miley sims|     2|   love|    night|
```

```
|The Endless Begin...|misc|Zecharia Sitchin|     5|   love|     life|
|The Endless Begin...|misc|Zecharia Sitchin|     5|   love|     like|
|The Endless Begin...|misc|Zecharia Sitchin|     5|   love|years ago|
|       Until The End| pop|             Six|     4|sadness|    alone|
|       Until The End| pop|             Six|     4|sadness|     fuck|
|       Until The End| pop|             Six|     4|sadness|     like|
|       Until The End| pop|             Six|     4|sadness|     shit|
|    Hasten the Blows| pop| Bluebottle Kiss|     8|  anger|     girl|
|    Hasten the Blows| pop| Bluebottle Kiss|     8|  anger|    heart|
```

Took 7 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:37:49 PM.

data-merge

```
%pyspark SPARK JOB (http://nyu-dataproc-sw-cfv5.c.hpc-dataproc-19b8.internal:46637/jobs/job?id=82)  FINISHED

output_dir = '/user/tw2770_nyu_edu/final-project/lyrics-emotion-keyword-rating'

result_df.write \
    .partitionBy("emotion", "keyword", "rating") \
    .mode("overwrite") \
    .option("compression", "snappy") \
    .parquet(output_dir)
```

Took 2 min 23 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:41:57 PM.