data-analytics

%pyspark SPARK JOB FINISHED

```
%pyspark

file_path = '/user/tw2770_nyu_edu/final-project/lyrics-emotion-rating'

df = spark.read.parquet(file_path)

df.show(20)
```

```
+-------+--------------------+---+----------------+--------------------+-------+------+
|     id|               title|tag|          artist|              lyrics|emotion|rating|
+-------+--------------------+---+----------------+--------------------+-------+------+
|1002094|          What I Say|pop|   The Yard Fight|I'm not trying to...|  anger|   3.0|
|1004616|     Whos In Season?|pop|        Ruination|They put a gun in...|  anger|   3.0|
|1007183|             Epílogo|pop|          Masacre|...bien llegó est...|  anger|   3.0|
|1008105|      Lukewarm Happy|pop| There Were Wires|If you could fath...|  anger|   3.0|
|1008109|               Rhino|pop|            Geist|Wieder ducke ich ...|  anger|   3.0|
|1008530|                  21|pop|        DeadWrong|Naive visions of ...|  anger|   3.0|
|1009582| Dark Is What I Want|pop|Apostle Of Hustle|Some rock, some t...|  anger|   3.0|
|1009800|           I Like It|pop|     Stephaniesid|Maybe i like when...|  anger|   3.0|
|1011439|             Culprit|pop|          Harmful|Culprit Pathetic....|  anger|   3.0|
|1011593|         Equilibrium|pop|            Nihil|I'm losing touch ...|  anger|   3.0|
|1011832|  Get Someplace Else|pop|   Blood Meridian|An empty road bes...|  anger|   3.0|
|1013066|  Clouds On The Moon|pop|Fire In The Skies|We are the same.T...|  anger|   3.0|
|1013681|One Cheque From t...|pop| Dayglo Abortions|There comes a tim...|  anger|   3.0|
|1014042|  Amor A Contrapunto|pop|     Daniela Romo|No soy lo que esp...|  anger|   3.0|
|1014116|       One More Time|pop|    Standing Tall|No control (tired...|  anger|   3.0|
```

Took 1 sec. Last updated by anonymous at December 10 2024, 2:00:10 AM.

Calculate the distribution of `rating` column

FINISHED

Took 2 sec. Last updated by anonymous at December 10 2024, 2:01:15 AM.

%pyspark SPARK JOB FINISHED

```
%pyspark

rating_distribution = df.groupBy("rating").agg(
    count("*").alias("count")
)

# Show the result
rating_distribution.orderBy("rating").show()
```

```
+------+-----+
|rating|count|
+------+-----+
|   1.0|  534|
|   2.0| 4154|
|   3.0| 8776|
|   4.0| 5167|
|   5.0| 1554|
|   6.0|  347|
|   7.0|   79|
|   8.0|   20|
|   9.0|    4|
|  10.0|    2|
+------+-----+
```

Took 1 sec. Last updated by anonymous at December 10 2024, 2:00:44 AM.

Calculate the distribution among the emotions

FINISHED

Took 0 sec. Last updated by anonymous at December 09 2024, 2:07:05 PM.

SPARK JOB FINISHED

```
%pyspark

from pyspark.sql.functions import col, round, avg, min, max, count, row_number
from pyspark.sql import Window

emotion_counts = df.groupBy("emotion").count()

# Calculate total number of records
total = df.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()
```

```
+--------+-----+----------+
| emotion|count|percentage|
+--------+-----+----------+
|   anger|10727|     51.98|
|     joy| 5579|     27.03|
| sadness| 2516|     12.19|
|    love|  862|      4.18|
|    fear|  813|      3.94|
|surprise|  140|      0.68|
+--------+-----+----------+
```

Took 1 sec. Last updated by anonymous at December 10 2024, 2:18:36 AM.

Analyze the relationship between `emotion` and `rating`

FINISHED

Took 0 sec. Last updated by anonymous at December 09 2024, 3:28:59 PM.

SPARK JOB FINISHED

```
%pyspark

emotion_rating_stats = df.groupBy("emotion").agg(
    avg("rating").alias("average"),
    min("rating").alias("min"),
    max("rating").alias("max")
)

emotion_rating_stats.show()
```

```
+--------+------------------+---+----+
| emotion|           average|min| max|
+--------+------------------+---+----+
|     joy|3.1398100017924357|1.0| 9.0|
|    love|3.2795823665893273|1.0| 9.0|
|    fear| 3.257072570725707|1.0| 8.0|
|   anger| 3.223827724433672|1.0|10.0|
| sadness|3.3569157392686804|1.0| 9.0|
|surprise|3.1357142857142857|1.0| 6.0|
+--------+------------------+---+----+
```

Took 1 sec. Last updated by anonymous at December 10 2024, 2:18:46 AM.

SPARK JOB FINISHED

```
%pyspark

# Group by rating and emotion, then count the occurrences
emotion_rating_distribution = df.groupBy("rating", "emotion").agg(
    count("*").alias("count")
).orderBy("rating", "emotion")

emotion_rating_distribution.show(60)
```

```
+------+--------+-----+
|rating| emotion|count|
+------+--------+-----+
|   1.0|   anger|  234|
|   1.0|    fear|   19|
|   1.0|     joy|  214|
|   1.0|    love|   28|
|   1.0| sadness|   36|
|   1.0|surprise|    3|
|   2.0|   anger| 2139|
|   2.0|    fear|  151|
|   2.0|     joy| 1325|
|   2.0|    love|  172|
|   2.0| sadness|  333|
|   2.0|surprise|   34|
|   3.0|   anger| 4672|
|   3.0|    fear|  342|
```

Took 1 sec. Last updated by anonymous at December 10 2024, 2:19:02 AM.

SPARK JOB FINISHED

```
%pyspark

keyword_path = '/user/tw2770_nyu_edu/final-project/lyrics-emotion-keyword-rating'
keywords_df = spark.read.parquet(keyword_path)
keywords_df.show(100)
```

```
+--------------------+----+--------------------+-------+-------+------+
|               title| tag|              artist|emotion|keyword|rating|
+--------------------+----+--------------------+-------+-------+------+
|           I Like It| pop|        Stephaniesid|  anger|   like|   3.0|
|         Equilibrium| pop|               Nihil|  anger|   like|   3.0|
|The Last Gate The...| pop|                Root|  anger|   like|   3.0|
|  Clouds On The Moon| pop|   Fire In The Skies|  anger|   like|   3.0|
|       One More Time| pop|       Standing Tall|  anger|   like|   3.0|
|     Went Like Lambs| pop|             Quitter|  anger|   like|   3.0|
|Fearless Vampire ...| pop|A Spectre Is Haun...|  anger|   like|   3.0|
|         A Situation| pop|      The Kudzu Wish|  anger|   like|   3.0|
|          Heads Up| pop|       Diehard Youth|  anger|   like|   3.0|
|      Now You See It| pop|        Jump Smokers|  anger|   like|   3.0|
|Sunken Pleasure /...| pop|The Legendary Pin...|  anger|   like|   3.0|
|              Raygun| pop|           Tolerance|  anger|   like|   3.0|
|    New England Rain| pop|       Matt McIntosh|  anger|   like|   3.0|
|      Stormy Weather| pop|        Jimmy Luxury|  anger|   like|   3.0|
```

Took 3 sec. Last updated by anonymous at December 10 2024, 2:41:22 AM.

Calculate occurences for each of the keyword

FINISHED

Took 0 sec. Last updated by anonymous at December 09 2024, 3:29:46 PM.

SPARK JOB FINISHED

```
%pyspark
keyword_count = keywords_df.groupBy("keyword").agg(
    count("*").alias("total_count")
)

keyword_count.orderBy(col("total_count").desc()).show()
```

```
+-------+-----------+
|keyword|total_count|
+-------+-----------+
|   like|       6749|
|   life|       3663|
|   fuck|       3162|
|   love|       3147|
|    die|       2682|
|  night|       2419|
|   mind|       2209|
|  heart|       2127|
|   shit|       2098|
|  world|       2087|
|   girl|       1535|
|  bitch|       1503|
|  leave|       1492|
|   pain|       1467|
|   stay|       1378|
```

Took 5 sec. Last updated by anonymous at December 10 2024, 2:41:33 AM.

Top 10 keywords

FINISHED

Took 0 sec. Last updated by anonymous at December 09 2024, 3:30:35 PM.

SPARK JOB FINISHED

```
%pyspark
top_10_keywords = keyword_count.orderBy(col("total_count").desc()).limit(10).drop("total_c

top_10_keywords.show()
```

```
+-------+
|keyword|
+-------+
|   like|
|   life|
|   fuck|
|   love|
|    die|
|  night|
|   mind|
|  heart|
|   shit|
|  world|
+-------+
```

Took 5 sec. Last updated by anonymous at December 10 2024, 2:41:49 AM.

Calculate the distribution among the emotions for each keyword

FINISHED

Took 0 sec. Last updated by anonymous at December 09 2024, 3:31:37 PM.

SPARK JOB FINISHED

```
%pyspark

keyword_emotion_count = keywords_df.groupBy("keyword", "emotion").agg(
    count("*").alias("count")
)

keyword_emotion_count.orderBy("keyword", col("count").desc()).show()
```

```
|   alone|surprise|    6|
|ay ay ay|   anger|    9|
|ay ay ay|     joy|    1|
|    baby|   anger|  450|
|    baby|     joy|  209|
|    baby|    love|  121|
|    baby| sadness|   96|
|    baby|    fear|   33|
|    baby|surprise|    9|
|    best|   anger|  577|
|    best|     joy|  457|
|    best| sadness|  161|
|    best|    love|   66|
|    best|    fear|   45|
|    best|surprise|   15|
+--------+--------+-----+
only showing top 20 rows
```

Took 5 sec. Last updated by anonymous at December 10 2024, 2:42:01 AM.

Show the distribution among the emotions of the top 10 keywords

FINISHED

Took 0 sec. Last updated by anonymous at December 09 2024, 3:32:24 PM.

SPARK JOB FINISHED

```
%pyspark

# Filter keyword_emotion_count for only the top 10 keywords
top_10_keyword_emotion_count = keyword_emotion_count.join(
    top_10_keywords, "keyword", "inner"
)

# Show the result
[illegible]
```

```
|  night| sadness|  389|
|  night|    fear|  134|
|  night|    love|  157|
|  night|surprise|   15|
|   shit|   anger| 1443|
|   shit|     joy|  324|
|   shit| sadness|  182|
|   shit|    fear|   57|
|   shit|    love|   73|
|   shit|surprise|   19|
|  world|   anger|  897|
|  world|     joy|  624|
|  world| sadness|  363|
|  world|    love|   96|
|  world|    fear|   98|
|  world|surprise|    9|
+-------+--------+-----+
```

Took 17 sec. Last updated by anonymous at December 10 2024, 2:42:24 AM.

Calculate the keyword-rating matrix

FINISHED

Took 3 sec. Last updated by tw2770_nyu_edu at December 10 2024, 4:05:27 PM.

SPARK JOB FINISHED

```
%pyspark
from pyspark.sql.functions import col, count, sum, round

# Group by keyword and rating to count occurrences
keyword_rating_counts = keywords_df.groupBy("keyword", "rating").agg(count("*").alias("cou

# Sum counts by keyword to calculate total counts per keyword
keyword_totals = keyword_rating_counts.groupBy("keyword").agg(sum("count").alias("total_co

# Join keyword_rating_counts with keyword_totals to calculate percentages
keyword_rating_percentages = keyword_rating_counts.join(
    keyword_totals, on="keyword"
).withColumn(
    "percentage", round((col("count") / col("total_count")), 2)
)

# Pivot table to create matrix
keyword_rating_matrix = keyword_rating_percentages.groupBy("keyword").pivot("rating").agg(
    sum("percentage")
).fillna(0)

# Show the resulting matrix
keyword_rating_matrix.show(45)
```

```
|    yo bitch| 0.0|0.24|0.46|0.17| 0.1|0.02|0.01|0.01|0.0|
| really want| 0.0|0.22|0.47|0.18|0.12|0.01| 0.0| 0.0|0.0|
|    new york| 0.0|0.18|0.29|0.35|0.18| 0.0| 0.0| 0.0|0.0|
|    na na na| 0.0| 0.2| 0.3| 0.3| 0.2| 0.0| 0.0| 0.0|0.0|
|never forget| 0.0|0.14|0.45|0.29| 0.1|0.02| 0.0| 0.0|0.0|
|   years ago|0.03|0.24|0.32|0.26| 0.1|0.03| 0.0|0.02|0.0|
|    ay ay ay| 0.0| 0.4| 0.4| 0.2| 0.0| 0.0| 0.0| 0.0|0.0|
|    say love| 0.0|0.25| 0.5|0.25| 0.0| 0.0| 0.0| 0.0|0.0|
|  never stop|0.02|0.13|0.42|0.29| 0.1|0.05| 0.0| 0.0|0.0|
+------------+----+----+----+----+----+----+----+----+---+
```

Took 12 sec. Last updated by anonymous at December 10 2024, 2:50:35 AM.

Calculate the keyword-genre matrix

FINISHED

Took 0 sec. Last updated by tw2770_nyu_edu at December 10 2024, 4:05:53 PM. (outdated)

SPARK JOB FINISHED

```
%pyspark
# Group by keyword and tag to count occurrences
keyword_tag_counts = keywords_df.groupBy("keyword", "tag").agg(count("*").alias("count"))

# Calculate total counts per keyword
keyword_totals = keyword_tag_counts.groupBy("keyword").agg(sum("count").alias("total_count

# Join to calculate percentages
keyword_tag_percentages = keyword_tag_counts.join(
    keyword_totals, on="keyword"
).withColumn(
    "percentage", round((col("count") / col("total_count")), 2)
)

# Pivot table to create matrix
keyword_tag_matrix = keyword_tag_percentages.groupBy("keyword").pivot("tag").agg(
    sum("percentage")
).fillna(0)

# Show the resulting matrix
keyword_tag_matrix.show(45)
```

```
|  never ever|    0.0|0.06|0.48|0.39|0.03|0.02|
|        fuck|    0.0|0.07|0.28|0.53|0.02|0.11|
| every night|   0.02|0.11|0.45|0.27|0.02|0.14|
|         die|    0.0|0.16| 0.4| 0.3|0.01|0.13|
|    oh oh oh|    0.0|0.16|0.47|0.25|0.13| 0.0|
|         god|    0.0|0.15|0.32| 0.4|0.01|0.12|
|        best|   0.01|0.23| 0.3|0.35|0.02|0.09|
|    yo bitch|    0.0|0.03|0.02|0.95| 0.0| 0.0|
| really want|   0.01| 0.1|0.27|0.55|0.01|0.06|
|    new york|    0.0|0.18|0.35|0.47| 0.0| 0.0|
|    na na na|    0.0| 0.1| 0.5| 0.2| 0.1| 0.1|
|never forget|    0.0| 0.1|0.51|0.21|0.01|0.17|
|   years ago|   0.02|0.32|0.43|0.16| 0.0|0.07|
|    ay ay ay|    0.1| 0.0| 0.3| 0.6| 0.0| 0.0|
|    say love|    0.0|0.13|0.38|0.38|0.13| 0.0|
|  never stop|   0.01|0.06| 0.4| 0.4| 0.0|0.12|
+------------+-------+----+----+----+----+----+
```

Took 14 sec. Last updated by anonymous at December 10 2024, 3:18:50 AM.