%spark.pyspark **SPARK JOB FINISHED KEYWORDS****SPARK JOB FINISHED

CHOOLING

CHOOLING

**SPARK JOB FINISHED

CHOOLING

**CHOO

df = spark.read.parquet(file_path)
df.show()

l title tag	artistlyearlvi	ews lyrics id r	ratingl
	Jhenny gonzaga 2015	0 be glad if I tell 2022005	 1l
Ur Stanton Nildab rock	Sam Asial2015l	0 We empower our cl 2147320	11
	PropLadder 2015	0 Provident prestig 2303729	11
l Jest 11miscl T	arlton's Jests 2015	0 To an ordinary in 2312437	11
	ırasaki Shikibul2022l	0 CharacterDescript 7564366	11
l Coś o mnielmiscl	Amadej WalczakI2022I	0 Po ukończeniu stu 7651058	11
l Geroganl raplDJ	Hamid Khareji 2021	0 In an interview, 6768076	11
l Golemirocki	Akimbo (Band) 2003	0 For three years n 6786967	11
Cavernous Black Eyes rock	Akimbo (Band) 2003	0 All held back fea 6786972	11
Skipping Stones pop	Horace Pinker 2000	0 When i looked acr 7705496	11
Carnival Nostalgia pop	Horace Pinker 2000	0 I laughed with hi 7705497	11
l If You FallIrock!	Horace Pinker 2005	0 If you fall on yo 7702041	11
l Half Waylrockl	Horace Pinker 2005	0 Hardly walking an 7702042	11
l Resonate So Reallrockl	Horace Pinker 2005	0 He never walked a 7702043	11
Dolinoi 101 I model	Hamasa Dinkam 200F	AlThomata no magula 177020221	1 1

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:50 PM.

%spark.pyspark df.printSchema() **FINISHED**

root

I-- title: string (nullable = true)
I-- tag: string (nullable = true)
I-- artist: string (nullable = true)
I-- year: integer (nullable = true)
I-- views: integer (nullable = true)
I-- lyrics: string (nullable = true)
I-- id: string (nullable = true)
I-- rating: integer (nullable = true)

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:50 PM.

Top Keywords

FINISHED

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

Romove punctuation of lyrics column

FINISHED

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:05:54 PM.

%spark.pyspark from pyspark.sql import SparkSession **FINISHED**

from pyspark.ml.feature import Tokenizer, NGram
from pyspark.sql.functions import explode, col, lower, regexp_replace

Remove punctuations and convert text to lowercase

keýwórdá-sélectionember 09 2024, 1:02:50 PM.

%spark.pyspark
df.printSchema()

root

I-- title: string (nullable = true)
I-- tag: string (nullable = true)
I-- artist: string (nullable = true)
I-- year: integer (nullable = true)
I-- views: integer (nullable = true)
I-- lyrics: string (nullable = true)
I-- id: string (nullable = true)
I-- rating: integer (nullable = true)

Took 1 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:51 PM.

%spark. **国**\$PARK JOB (http://nyu-dataproc-sw-cfv5.c.hpc-dataproc-19b8.internal:46637/jobs/job?id=19) FINISHED df.show()

+		++	+	+-	+
l titlel tagl	artistlyear	lviewsl	lyrics	idlr	ratingl
+	+	++	+	+-	+
I TO LOVE YOUImiscl Jhen	ny gonzaga 2 <mark>01</mark> 5	l 01	be glad if i tell	20220051	11
Ur Stanton Nildab rock	Sam Asia 2015	l 01	we empower our cl	21473201	11
Provident Kenwort misc	PropLadder 2015	l 01	provident prestig	23037291	11
Jest 1 misc Tarlt	on's Jests 2015	l 01	to an ordinary in	23124371	11
Tale of Genji Cha misc Murasa	ki Shikibu 2022	l 01	characterdescript	75643661	11
l Coś o mnielmiscl Amad	ej Walczak 2 <mark>0</mark> 22	l 01	po ukoczeniu stud	76510581	11
l Gerogani rapiDJ Ham	id Khareji 2021	l 01	in an interview h	67680761	11
Golemirock Aki	mbo (Band) 2003	l 01	for three years nl	67869671	11
	mbo (Band) 2003	l 01	all held back fea	67869721	11
<pre>Skipping Stones pop Hore</pre>	ace Pinker 2000	l 01	when i looked acr	77054961	11
l Carnival Nostalgial popl Hor	ace Pinker 2000	l 01	i laughed with hi	77054971	11
I If You FallIrockI Hor	ace Pinker 2005	l 01	if you fall on yo	7702041	11
l Half Waylrockl Hor	ace Pinker 2005	l 01	hardly walking an	77020421	11
l Resonate So Realirocki Hor	ace Pinker 2 <mark>00</mark> 5	l 01	he never walked a	77020431	11
1 B 3' '4041 11 11	D' 1 1300E		• • • • •	77020221	4.1

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:51 PM.

Split lyrics column into tokens

FINISHED

FINISHED

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:51 PM.

```
%spark.pyspark
# Tokenize the lyrics into individual words
tokenizer = Tokenizer(inputCol="lyrics", outputCol="words")
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:51 PM.

words_data = tokenizer.transform(df).persist()

localhost:37479/#/notebook/2KE3JS4EK

%spark. **国\$PXT**KK JOB (http://nyu-dataproc-sw-cfv5.c.hpc-dataproc-19b8.internal:46637/jobs/job?id=20) FINISHED words_data.show()

_			
keywords selectio	DUIII ASTUIZATOI	שושe empower our כנובד+ו אבושר	ΤI
	PropLadder 2015	Olprovident prestig 2303729	11
[provident, prest			
Jest 1 misc Ta	rlton's Jests 2015	0lto an ordinary in 2312437	11
[to, an, ordinary		•	
	asaki Shikibu 2022	0 characterdescript 7564366	11
[characterdescrip		·	
l Coś o mnielmiscl A	madej Walczak 2022	0lpo ukoczeniu stud 7651058	11
[po, ukoczeniu, s	·	·	
_, ,	Hamid Khareji 2021	0lin an interview h 6768076	11
[in, an, intervie	j		
- · ·	Akimbo (Band) 2003	Olfor three years n 6786967	11
[for, three, year	, ,	·	
	Akimbo (Band) 2003	0 all held back fea 6786972	11
[all, held, back,			
_ , , ,	Horace Pinker 2000	0 when i looked acr 7705496	11
[when, i, looked,			
<u> </u>			
Took 0 sec. Last updated by tw2770_nyu_edu at I	December 09 2024, 1:02:51 PM.		

Unigram

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:05:43 PM.

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:51 PM.

```
#spark.pyspark
from pyspark.ml.feature import StopWordsRemover, Tokenizer
from pyspark.sql.functions import col, explode, concat_ws

default_stop_words = StopWordsRemover.loadDefaultStopWords("english")
unigram_stop_words = [" ", 'im', 'dont', 'know', 'got', 'get', 'time', 'one', 'cant', 's
    'think', 'let', 'man', 'que', 'back', 'thats', 'feel', 'cause', 'still', 'day', 'away'
    'nigga']
extend_uni_stop_words = default_stop_words + unigram_stop_words
```

```
%spark.pyspark FINISHED
```

['i', 'me', 'my', 'myself', 'we', 'our', 'ours', 'ourselves', 'you', 'your', 'yours', 'your self', 'yourselves', 'he', 'him', 'his', 'himself', 'she', 'her', 'hers', 'herself', 'it', 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'a nd', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'be low', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'furt her', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', 'should', 'no w', "i'll", "you'll", "he'll", "she'll", "they'll", "i'd", "you'd", "he'd", "she'd", "we'd", "they'd", "i'm", "you're", "he's", "she's", "it's", "we're", "they're", "i'v

print(default_stop_words)

e", "we've", "you've", "they've", "isn't", "aren't", "wasn't", "weren't", "haven't", "has n't", "hadn't", "doesn't", "didn't", "won't", "wouldn't", "shan't", "shan't", "shouldn't", "mustn't", "can't", "couldn't", 'cannot', 'could', "here's", "how's", "let's", 'ought', "the stic", "wheels", "whee

keywords discretion ember 09 2024, 1:02:52 PM.

%spark.pyspark
print(extend_uni_stop_words)

FINISHED

['i', 'me', 'my', 'myself', 'we', 'our', 'ourselves', 'you', 'your', 'yours', 'your self', 'yourselves', 'he', 'him', 'his', 'himself', 'she', 'her', 'hers', 'herself', 'it', 'its', 'itself', 'they', 'them', 'their', 'theirs', 'themselves', 'what', 'which', 'who', 'whom', 'this', 'that', 'these', 'those', 'am', 'is', 'are', 'was', 'were', 'be', 'been', 'being', 'have', 'has', 'had', 'having', 'do', 'does', 'did', 'doing', 'a', 'an', 'the', 'a nd', 'but', 'if', 'or', 'because', 'as', 'until', 'while', 'of', 'at', 'by', 'for', 'with', 'about', 'against', 'between', 'into', 'through', 'during', 'before', 'after', 'above', 'be low', 'to', 'from', 'up', 'down', 'in', 'out', 'on', 'off', 'over', 'under', 'again', 'furt her', 'then', 'once', 'here', 'there', 'when', 'where', 'why', 'how', 'all', 'any', 'both', 'each', 'few', 'more', 'most', 'other', 'some', 'such', 'no', 'nor', 'not', 'only', 'own', 'same', 'so', 'than', 'too', 'very', 's', 't', 'can', 'will', 'just', 'don', 'should', 'no w', "i'll", "you'll", "he'll", "she'll", "they'll", "i'd", "you'd", "he'd", "she'd", "we'd", "they'd", "i'm", "you're", "he's", "she's", "it's", "we're", "they're", "i'v e", "we've", "you've", "they've", "isn't", "aren't", "wasn't", "weren't", "haan't", "haon't", "don't", "didn't", "won't", "wouldn't", "shouldn't", "mustn't", "can't", "couldn't", 'cannot', 'could', "here's", "how's", "let's", 'ought', "that's", "there's", "what's", "where's", "who's", "why's", 'would', ' ', 'im', 'd

Took 0 sec. Last updated by tw2770 nvu edu at December 09 2024. 1:02:52 PM.

```
%spark.pyspark
# Remove stop words using the custom list
remover = StopWordsRemover(inputCol="words", outputCol="filtered_words")
remover.setStopWords(extend_uni_stop_words)
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:52 PM.

unigram_data = remover.transform(words_data)

```
%spark. 摩$PARK JOB (http://nyu-dataproc-sw-cfv5.c.hpc-dataproc-19b8.internal:46637/jobs/job?id=21) FINISHED unigram_data.show()
```

```
IUI STUITUII NELUUD...IIOCKI
                                                       שושe empower. our. Cl...ועבאראכעו
                                   JUII AS LU I ZU J I
[we, empower, our...|[empower, clients...|
| IProvident Kenwort...|miscl
                                  PropLadder|2015|
                                                       Olprovident prestig... | 2303729|
                                                                                            11
[provident, prest...|[provident, prest...|
               Jest 1|misc| Tarlton's Jests|2015|
                                                       Olto an ordinary in... | 2312437 |
                                                                                            11
[to, an, ordinary...|[ordinary, whitef...|
| ITale of Genji Cha...|misc|Murasaki Shikibu|2022|
                                                       Olcharacterdescript...|7564366|
                                                                                            11
[characterdescrip...|[characterdescrip...|
           Coś o mnielmiscl Amadej Walczakl20221
                                                       Olpo ukoczeniu stud...|7651058|
                                                                                            11
[po, ukoczeniu, s...| [po, ukoczeniu, s...|
              Geroganl rapIDJ Hamid Kharejil20211
                                                       Olin an interview h... | 6768076|
                                                                                            11
[in, an, intervie... | [interview, hamid... |
                Golemirocki
                              Akimbo (Band)|2003|
                                                       Olfor three years n... | 6786967|
                                                                                            11
[for, three, year...|[three, years, st...|
|Cavernous Black Eyes|rock| Akimbo (Band)|2003|
                                                       0|all held back fea...|6786972|
                                                                                            11
[all, held, back,...|[held, fearing, d...|
                                                       Olwhen i looked acr... | 7705496 |
      Skipping Stones | popl
                              Horace Pinker|2000|
                                                                                            11
[when, i, looked,...|[looked, across, ...|
```

```
%spark.pyspark ■ SPARK JOB FINISHED
```

```
unigrams_flattened = unigram_data.select(explode(col("filtered_words")).alias("unigram"))
unigram_counts = unigrams_flattened.groupBy("unigram").count()
unigram_counts_sorted = unigram_counts.orderBy(col("count").desc())
unigram_counts_sorted.show(100)
```

```
placel 16841
      playl 16831
      non| 1674|
      showl
            16611
      painl 16591
      chel 16491
      lost| 1603|
     alonel 15801
     thingl 15731
   without | 1572|
     wordsl 15571
      hearl
            1551 l
        dal 15501
      name | 1544 |
+----+
only showing top 100 rows
```

Took 3 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:55 PM.

```
%spark.pyspark FINISHED
```

```
uni_words = ['like', 'life', 'love', 'yeah', 'shit', 'fuck', 'bitch', 'world', 'mind', 'he
    , 'stay', 'night', 'best']
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:55 PM.

Bigram

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:06:26 PM.

```
%spark.pyspark FINISHED
```

```
bigram_stop_words = [" ", 'im', 'dont', 'know', 'got', 'get', 'time', 'one', 'cant', 'se
   'think', 'let', 'man', 'que', 'back', 'thats', 'feel', 'cause', 'still', 'day', 'away'
   'interview', 'nigga', 'looks', 'look']
extend_bi_stop_words = default_stop_words + bigram_stop_words
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:55 PM.

%spark. **国**\$PARK JOB (http://nyu-dataproc-sw-cfv5.c.hpc-dataproc-19b8.internal:46637/jobs/job?id=24) FINISHED words_data.show()

9/24, 1:15 PM	keywords-selection - Zeppelin		
I JUSE TIMESEI	INI LUNI 3 JUSUSILUTJI	OICO UII OIULIIUI Y LIIILJILTJI I	± 1
[to, an, ordinary			
ITale of Genji ChalmiscIMu	urasaki Shikibu 2022	0 characterdescript 7564366	11
[characterdescrip			
CDÉ o mnielmisch	Amadej Walczak 2022	0lpo ukoczeniu stud 7651058	11
keywords-selecti	OH		
	J Hamid Khareji 2021	0lin an interview h 6768076	1!
[in, an, intervie	-		
l Golemirocki	Akimbo (Band) 2003	0lfor three years n 6786967	11
[for, three, year			
Cavernous Black Eyes rock	Akimbo (Band) 2003	0 all held back fea 6786972	11
[all, held, back,			
<pre>Skipping Stones pop </pre>	Horace Pinker 2000	0 when i looked acr 7705496	11
Twhen i looked I			
Took 0 sec. Last updated by tw2770_nyu_edu	at December 09 2024, 1:02:55 PM.		

```
%spark.pyspark
# Remove stop words using the custom list
remover = StopWordsRemover(inputCol="words", outputCol="filtered_words")
remover.setStopWords(extend_bi_stop_words)
bigrams_data = remover.transform(words_data)
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:55 PM.

```
%spark. 國知林 JOB (http://nyu-dataproc-sw-cfv5.c.hpc-dataproc-19b8.internal:46637/jobs/job?id=25) FINISHED
bigrams_data.show()
IUI. STALLOLI METAAD...ILOCKI
                                     JUII AS LU I ZU ZU I
                                                        שושe empower our כנ...ובבאר שוש
[we, empower, our...|[empower, clients...|
| IProvident Kenwort...|miscl
                                  PropLadder | 2015 |
                                                        Olprovident prestig... | 2303729 |
                                                                                              11
[provident, prest...|[provident, prest...|
               Jest 1|misc| Tarlton's Jests|2015|
                                                        Olto an ordinary in...|2312437|
                                                                                              11
[to, an, ordinary...|[ordinary, whitef...|
| Tale of Genji Cha...| misc| Murasaki Shikibu| 2022 |
                                                        0|characterdescript...|7564366|
                                                                                              11
[characterdescrip...|[characterdescrip...|
           Coś o mnielmiscl Amadej Walczakl20221
                                                        Olpo ukoczeniu stud...|7651058|
                                                                                              11
[po, ukoczeniu, s...|[po, ukoczeniu, s...|
              Geroganl rapIDJ Hamid Kharejil20211
                                                        Olin an interview h... | 6768076|
                                                                                              1 I
[in, an, intervie...|[hamid, khareji, ...|
                Golemirocki
                               Akimbo (Band)|2003|
                                                        Olfor three years n... | 6786967|
                                                                                              11
[for, three, year...|[three, years, st...|
                                                        0|all held back fea...|6786972|
| | Cavernous Black Eyes|rock|
                               Akimbo (Band)|2003|
                                                                                              11
[all, held, back,...|[held, fearing, d...|
      Skipping Stones | popl
                                                        Olwhen i looked acr... 177054961
                                                                                              11
                               Horace Pinker | 2000 |
[when, i, looked,...|[looked, across, ...|
```

```
%spark.pyspark
bigram = NGram(n=2, inputCol="filtered_words", outputCol="bigrams")
bigram_data = bigram.transform(bigrams_data)
FINISHED
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:56 PM.

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:55 PM.

%spark. **医\$PATR**K JOB (http://nyu-dataproc-sw-cfv5.c.hpc-dataproc-19b8.internal:46637/jobs/job?jd=26) FINISHED bigram_data.show()

```
ITTOVERCITE RETINOTE...TIMESET
                                   110pEuuuci 120131
                                                         [provident, prest...|[provident, prest...|[provident presti...|
                Jest 1|misc| Tarlton's Jests|2015|
                                                         Olto an ordinary in... | 2312437 |
                                                                                                1 I
[to, an, ordinary...|[ordinary, whitef...|[ordinary whitefr...|
Tale of Genji Cha. Imisc Murasaki Shikibul20221     0|characte
Bance des ipse [enclude descrip...|[characterdescrip...|
                                                         Olcharacterdescript...175643661
                                                                                                11
           Coś o mnielmisc! Amadej Walczak!2022!
                                                         Olpo ukoczeniu stud...|7651058|
                                                                                                11
[po, ukoczeniu, s...|[po, ukoczeniu, s...|[po ukoczeniu, uk...|
               Geroganl rapIDJ Hamid Kharejil20211
                                                         Olin an interview h... | 6768076|
                                                                                                1 I
[in, an, intervie...|[hamid, khareji, ...|[hamid khareji, k...|
                                Akimbo (Band)|2003|
                 Golemirocki
                                                         Olfor three years n... | 6786967 |
                                                                                                11
[for, three, year...|[three, years, st...|[three years, yea...|
| | Cavernous Black Eyes|rock|
                                Akimbo (Band)|2003|
                                                         0|a|l held back fea...|6786972|
                                                                                                11
[all, held, back,...|[held, fearing, d...|[held fearing, fe...|
      Skipping Stones | pop | Horace Pinker | 2000 |
                                                         Olwhen i looked acr... | 7705496 |
                                                                                                11
Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:02:56 PM.
```

```
%spark.pyspark
                                                                                  FINISHED
# Explode the bigrams list to get individual bigrams as rows
bigrams_flattened = bigram_data.select(explode(col("bigrams")).alias("bigram"))
# Calculate frequency by grouping by bigram and counting the occurrences
bigram_counts = bigrams_flattened.groupBy("bigram").count()
# Sort the bigrams in descending order based on count
bigram_counts_sorted = bigram_counts.orderBy(col("count").desc())
```

Took 0 sec. Last updated by tw2770 nyu edu at December 09 2024. 1:02:56 PM.

```
%spark.pyspark
 bigram_counts_sorted.show(100)
          waririai
                    TTOI
         non mil
                    1181
                    1181
    leave alone!
        verse 21
                    117 l
      suck dickl
                    117 l
     never stopl
                    1171
Ι
   never forgetl
                    115 l
                    114|
             y yl
                    1131
             sayl
                    113 l
    never neverl
                    1121
      years agol
     dead girls!
                    1111
       never gol
                    1111
                    1111
          want l
    left behindl
                    1101
only showing top 100 rows
```

Took 7 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:03:03 PM.

```
%spark.pyspark
bi_words = ['every night', 'need loving', 'nothing left', 'yo bitch', 'say goodbye', 'best
    'say love', 'leave alone', 'never stop', 'never forget', 'years ago']
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:03:03 PM.

SPARK JOB FINISHED

Trigram

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:06:49 PM.

keywords-selection

```
%spark.pyspark FiNiSHED
```

```
trigram_stop_words = [" ", 'im', 'dont', 'know', 'got', 'get', 'time', 'one', 'cant', 's
    'think', 'let', 'man', 'que', 'back', 'thats', 'feel', 'cause', 'still', 'day', 'away'
    'interview', 'nigga', 'looks', 'look']
extend_tri_stop_words = default_stop_words + trigram_stop_words
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:03:03 PM.

```
%spark.pyspark FINISHED
```

```
# Remove stop words using the custom list
remover = StopWordsRemover(inputCol="words", outputCol="filtered_words")
remover.setStopWords(extend_tri_stop_words)
trigrams_data = remover.transform(words_data)
```

Took 1 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:03:04 PM.

```
%spark.摩软和K JOB (http://nyu-dataproc-sw-cfv5.c.hpc-dataproc-19b8.internal:46637/jobs/job?id=29) FINISHED trigrams_data.show()
```

```
IOI. SCALICOLI METARAD...ILOCKI
                                   JUII ASLUIZUIJI
                                                      שושe empower our כנ...ובבאר שוש
[we, empower, our...|[empower, clients...|
| IProvident Kenwort...|miscl
                                 PropLadder|2015|
                                                      Olprovident prestig...|2303729|
                                                                                            11
[provident, prest...|[provident, prest...|
               Jest 1|misc| Tarlton's Jests|2015|
                                                      Olto an ordinary in...|2312437|
                                                                                            11
ı
[to, an, ordinary...|[ordinary, whitef...|
| Tale of Genji Cha...|misc|Murasaki Shikibu|2022|
                                                      0|characterdescript...|7564366|
                                                                                            11
[characterdescrip...|[characterdescrip...|
          Coś o mnielmiscl Amadej Walczakl20221
                                                      Olpo ukoczeniu stud...|7651058|
                                                                                            11
[po, ukoczeniu, s...|[po, ukoczeniu, s...|
                                                                                            11
              Geroganl rapIDJ Hamid Kharejil20211
                                                      Olin an interview h... | 6768076 |
[in, an, intervie...|[hamid, khareji, ...|
                Golemirocki Akimbo (Band) | 2003 |
                                                      Olfor three years n... | 6786967|
                                                                                            11
[for, three, year...|[three, years, st...|
                                                      0|all held back fea...|6786972|
| | Cavernous Black Eyes|rock| Akimbo (Band)|2003|
                                                                                            11
[all, held, back,...|[held, fearing, d...|
      Skipping Stones | popl
                              Horace Pinker | 2000 |
                                                      Olwhen i looked acr... 177054961
                                                                                            11
[when, i, looked,...|[looked, across, ...|
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:03:04 PM.

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:03:04 PM.

```
%spark.pyspark
# Create trigrams using NGram
trigram = NGram(n=3, inputCol="filtered_words", outputCol="trigrams")
```

trigram_data = trigram.transform(trigrams_data)

%spark.**pspark**(JOB (http://nyu-dataproc-sw-cfv5.c.hpc-dataproc-19b8.internal:46637/jobs/job?id=30) FINISHED trigram_data.show()

```
+----+
     -----+
               title| tag|
                                 artistlyearlyiewsl
                                                             lyrics| id|rating|
                                trigramsl
 wordsl
            filtered_wordsl.
keywords-selection------
                          -----+
      TO LOVE YOU.....lmiscl Jhenny gonzagal2015
                                               0|be glad if i tell...|2022005|
  [be, glad, if, i,...|[glad, tell, miss...|[glad tell miss, ...|
  | IUr Stanton Nildab...|rock| | Sam Asia|2015|
                                               0 we empower our cl... | 2147320 |
                                                                               11
  [we, empower, our...|[empower, clients...|[empower clients ...|
  | Provident Kenwort...|misc| PropLadder|2015|
                                               Olprovident prestig...|2303729|
                                                                               1 I
  [provident, prest...|[provident, prest...|[provident presti...|
              Jest 1|misc| Tarlton's Jests|2015| Olto an ordinary in...|2312437|
                                                                               11
 [to, an, ordinary...|[ordinary, whitef...|[ordinary whitefr...|
  | Tale of Genji Cha...|misc|Murasaki Shikibu|2022|
                                               0|characterdescript...|7564366|
                                                                               11
  [characterdescrip...|[characterdescrip...|[characterdescrip...|
           Coś o mnielmisc| Amadej Walczak|2022| 0|po ukoczeniu stud...|7651058|
                                                                               11
                   I Fina intra carantina de la Fina intra carantina de la
     uka sa sa iu s
 Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:03:04 PM.
  %spark.pyspark
                                                                         FINISHED
  # Explode the trigrams list to get individual trigrams as rows
  trigrams_flattened = trigram_data.select(explode(col("trigrams")).alias("trigram"))
  # Calculate frequency by grouping by trigram and counting the occurrences
  trigram_counts = trigrams_flattened.groupBy("trigram").count()
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:03:04 PM.

Sort the trigrams in descending order based on count

trigram_counts_sorted = trigram_counts.orderBy(col("count").desc())

```
%spark.pyspark
                                                                            ■ SPARK JOB FINISHED
trigram_counts_sorted.show(100)
ICHILLEH CHILLEH C...I
                            <del>4</del>21
         hey hey heyl
                            441
                 neverl
                            441
           2021 2021
                            431
                right |
                            431
                 rightl
                            421
                            421
                  girll
                  20201
                            421
         work da workl
                            401
             ha ha hal
                            401
           da work dal
                            401
     peuttre auraitill
                            39 I
        dans une fortl
                            391
    diabolique moins!
                            391
 renfermer un secretl
                            391
        rsonne le mall
                            39 I
  tnbres locarina nal
                            39 I
       se rveille les!
                            39 I
```

```
%spark.pyspark
tri_words = ['juicy dead girls', 'ay ay ay', 'oh oh oh', 'na na na']
```

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:03:12 PM.

Took 8 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:03:12 PM.

Top Keywords

keywords-selection like, life, love, yeah, mind, heart, girl, die, baby, god, leave, best friend, say goodbye, really want, say love, leave alone, never stop, never forget

- 2. Hardship and Negativity: shit, fuck, bitch, world, pain, alone, nothing left, yo bitch, years ago
- 3. Lifestyle and Culture: money, stay, night, every night, need loving, new york, brand new, hip hop, juicy dead girls, ay ay ay, oh oh oh, na na na

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

%spark.pyspark FINISHED

keywords = uni_words + bi_words + tri_words
print(keywords)

['like', 'life', 'love', 'yeah', 'shit', 'fuck', 'bitch', 'world', 'mind', 'heart', 'girl', 'die', 'money', 'baby', 'god', 'leave', 'best', 'alone', 'pain', 'stay', 'night', 'best', 'every night', 'need loving', 'nothing left', 'yo bitch', 'say goodbye', 'best friend', 'ne ver ever', 'really want', 'new york', 'brand new', 'hip hop', 'say love', 'leave alone', 'n ever stop', 'never forget', 'years ago', 'juicy dead girls', 'ay ay ay', 'oh oh oh', 'na na na']

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:03:13 PM.

%spark.pyspark len(keywords) **FINISHED**

42

Took 0 sec. Last updated by tw2770_nyu_edu at December 09 2024, 1:03:13 PM.

%spark.pyspark READY