BDA Lab4

Kunal Sanjay Patil

20190802025

```
!pip install pyspark
```

```
    Requirement already satisfied: pyspark in /usr/local/lib/python3.7/dist-packages (3.2
    Requirement already satisfied: py4j==0.10.9.3 in /usr/local/lib/python3.7/dist-packag
```

```
from pyspark import SparkContext, SparkConf
sc = SparkContext.getOrCreate()
```

```
data = sc.textFile("books.csv")
```

```
type(data)
```

```
    pyspark.rdd.RDD
```

```
data.top(2)
```

```
    ['id,book_id,best_book_id,work_id,books_count,isbn,isbn13,authors,original_publicatic
     '9999,8565083,8565083,13433613,7,61711527,9.78006171153e+12,Peggy Orenstein,2011.0,(
```

```
data.collect()
```

```
     944,10584,10584,14015,152,451188462,9.78045118846e+12,Stephen King,1996.0,Desper
     '945,9369720,9369720,10958266,66,385739168,9.78038573916e+12,Lauren Kate,2011.0,P
     '946,13089710,13089710,18261314,66,385742886,9.78038574289e+12,James Dashner,2012
     '947,11500217,15863832,16435765,67,144477851X,9.78144477852e+12,Susan Ee,2011.0,A
     '948,187020,187020,826474,49,375726403,9.78037572641e+12,Richard Russo,2001.0,Emp
     '949,13895,13895,588034,56,1857232097,9.7818572321e+12,Robert Jordan,1993.0,The F
     "950,20820994,20820994,11409817,71,803734964,9.78080373497e+12,Jandy Nelson,2014.
     '951,22283,7076703,1602338,63,009928264X,9.78009928265e+12,Chuck Palahniuk,1999.0
     '952,9420,9420,3284660,80,385338708,9.78038533871e+12,Sophie Kinsella,2007.0,Shop
     '953,5091,5091,6309701,105,1416524525,9.78141652453e+12,Stephen King,2004.0,The D
     "954,6202690,6202690,6383190,14,316070521,9.78031607052e+12,Catherine Hardwicke,2
     '955,9418,9418,3238753,93,440241812,9.78044024181e+12,Sophie Kinsella,2001.0,Shop
     '956,5544,5544,321174,97,393316041,9.78039331605e+12,Richard Feynman,1984.0,"Sure
     '957,6654313,6654313,6848948,76,545123283,9.78054512328e+12,Maggie Stiefvater,201
     '958,3579,3579,7365,40,553609416,76783609419.0,L.M. Montgomery,1908.0,The Complet
     '959,7588,7588,3298883,766,142437344,9.78014243735e+12,"James Joyce, Seamus Deane
     '960,1166599,1166599,1920889,2,765302306,9.7807653023e+12,"Robert Jordan, Brandon
     '961,19089,19089,1461747,665,451529170,9.78045152918e+12,"George Eliot, Michel Fa
     '962,13496084,19390926,19038910,30,,,Colleen Hoover,2012.0,Point of Retreat,"Poir
     '963,5350,5350,1110983,118,385339100,9.7803853391e+12,John Grisham,1997.0,The Par
     '964,30,30,89369,74,345538374,9.78034553838e+12,J.R.R. Tolkien,1973.0,The Hobbit
     '965,13517535,13517535,14321512,38,146109111X,9.78146109111e+12,S.C. Stephens,200
     '966,425029,425029,7732,110,446350982,9.78044635098e+12,Scott Turow,1986.0,Presum
```

```
'967,11597,11597,1316297,131,045052468X,9.78045052468e+12,Stephen King,1989.0,The
'968,7968243,7968243,12020129,55,316001929,9.78031600192e+12,Stacy Schiff,2006.0,
'969,1911,1911,711993,83,374292795,9.7803742928e+12,Thomas L. Friedman,2005.0,The
'970,17182126,17182126,21366540,64,385743564,9.78038574356e+12,Brandon Sanderson,
'971,766020,766020,56647,75,1558580093,9.78155858009e+12,"Marcus Pfister, J. Alis
'972,32829,32829,1924715,1363,553213970,9.78055321397e+12,Jules Verne,1864.0,Voya
'973,6398634,6398634,6587328,73,61583251,9.78006158325e+12,Gretchen Rubin,2009.0,
'974,216363,216363,2398287,196,679740678,9.78067974067e+12,Philip K. Dick,1962.0,
'975,105992,105992,1077715,52,393322238,9.78039332223e+12,"Vincent Bugliosi, Curt
'976,90072,90072,24501333,3,793551617,9.78079355161e+12,"Robert Kapilow, Dr. Seus
'977,15645,15645,2377563,856,812970063,9.78081297007e+12,"Dante Alighieri, Anthor
'978,11559200,11559200,16499524,64,670023485,9.78067002349e+12,Deborah Harkness,2
'979,18816603,18816603,26757264,64,345544927,9.78034554493e+12,Jodi Picoult,2014.
'980,71811,71811,69530,48,441013813,9.78044101381e+12,Patricia Briggs,2006.0,Moor
'981,872333,872333,857698,47,786838922,9.78078683893e+12,Melissa de la Cruz,2006.
'982,40024,40024,2266643,70,812976142,9.78081297614e+12,Caleb Carr,1994.0,The Ali
'983,25489625,25489625,44848425,36,812993543,9.78081299355e+12,Ta-Nehisi Coates,2
'984,15997,15997,1031493,819,140424393,9.78014042439e+12,"John Milton, John
'985,6463967,26889576,6654434,57,393072231,9.78039307224e+12,Michael   Lewis,2009
'986,11564,11564,1836389,115,1416524290,9.7814165243e+12,Stephen King,1999.0,The
'987,8621462,8621462,13492114,138,1406311529,9.78140631152e+12,"Patrick Ness, Jim
'988,8253920,8253920,7338128,39,765325942,9.78076532594e+12,"Robert Jordan, Brand
'989,540020,540020,1792180,170,553266306,9.78055326631e+12,Frederick Forsyth,1971
'990,6732019,6732019,6928276,60,307463745,9.78030746375e+12,"Jason Fried, David H
'991,8909152,8909152,13785503,44,525951989,9.78052595199e+12,Rainbow Rowell,2011.
'992,3090465,3090465,6440505,36,739352350,9.78073935236e+12,"Stephenie Meyer, Ily
"993,12875258,12875258,18028067,49,679644199,9.78067964419e+12,Carol Rifka Brunt,
'994,32085,18929854,2247074,77,312965788,9.78031296578e+12,James Herriot,1972.0,A
'995,11408650,11408650,13460686,42,,,Michelle Hodkin,2011.0,The Unbecoming of Mar
'996,6854,6854,1341652,67,312966091,9.7803129661e+12,Janet Evanovich,1997.0,Three
'997,136116,136116,750426,427,1576469239,9.78157646923e+12,Emmuska Orczy,1905.0,T
'998,44186,44186,640276,23,037582913X,9.78037582913e+12,"Jon Stone, Michael J. Sm
'999,37741,37741,879886,66,425193799,9.78042519379e+12,Judy Blume,1972.0,Tales of
...]
```

```
data.take(2)
```

```
['id,book_id,best_book_id,work_id,books_count,isbn,isbn13,authors,original_publicatic
 '1,2767052,2767052,2792775,272,439023483,9.78043902348e+12,Suzanne Collins,2008.0,Th
```

```
for line in data.take(5):
 print(line)
```

```
id,book_id,best_book_id,work_id,books_count,isbn,isbn13,authors,original_publication_
1,2767052,2767052,2792775,272,439023483,9.78043902348e+12,Suzanne Collins,2008.0,The
2,3,3,4640799,491,439554934,9.78043955493e+12,"J.K. Rowling, Mary GrandPré",1997.0,Ha
3,41865,41865,3212258,226,316015849,9.78031601584e+12,Stephenie Meyer,2005.0,Twiligh
4,2657,2657,3275794,487,61120081,9.78006112008e+12,Harper Lee,1960.0,To Kill a Mockir
```

```
data.first()
```

```
     'id book id best book id work id books count isbn isbn13 authors original publicatio
```

```
oneRecord = data.first()
columns = oneRecord.split(',')
columns
```

```
     ['id',
      'book_id',
      'best_book_id',
      'work_id',
      'books_count',
      'isbn',
      'isbn13',
      'authors',
      'original_publication_year',
      'original_title',
      'title',
      'language_code',
      'average_rating',
      'ratings_count',
      'work_ratings_count',
      'work_text_reviews_count',
      'ratings_1',
      'ratings_2',
      'ratings_3',
      'ratings_4',
      'ratings_5',
      'image_url',
      'small_image_url']
```

```
import pyspark
from pyspark.sql import *
```

```
from pyspark.sql import SparkSession

spark = SparkSession.builder\
        .master("local")\
        .appName("Colab")\
        .config('spark.ui.port', '4050')\
        .getOrCreate()
```

```
type(spark)
```

```
     pyspark.sql.session.SparkSession
```

```
books_df = spark.read.csv('books.csv', header=True, inferSchema=True)
```

```
books_df.printSchema()
```

```
     root
      |-- id: integer (nullable = true)
      |-- book_id: integer (nullable = true)
      |-- best_book_id: integer (nullable = true)
      |-- work_id: integer (nullable = true)
      |-- books_count: integer (nullable = true)
      |-- isbn: string (nullable = true)
      |-- isbn13: double (nullable = true)
      |-- authors: string (nullable = true)
      |-- original_publication_year: double (nullable = true)
      |-- original_title: string (nullable = true)
      |-- title: string (nullable = true)
      |-- language_code: string (nullable = true)
      |-- average_rating: string (nullable = true)
      |-- ratings_count: string (nullable = true)
      |-- work_ratings_count: string (nullable = true)
      |-- work_text_reviews_count: string (nullable = true)
      |-- ratings_1: double (nullable = true)
      |-- ratings_2: integer (nullable = true)
      |-- ratings_3: integer (nullable = true)
      |-- ratings_4: integer (nullable = true)
      |-- ratings_5: integer (nullable = true)
      |-- image_url: string (nullable = true)
      |-- small_image_url: string (nullable = true)
```

```
type(books_df)
```

```
     pyspark.sql.dataframe.DataFrame
```

```
len(books_df.columns)
```

```
     23
```

```
ratings_df = spark.read.csv('/content/ratings.csv', header=True, inferSchema=True)
```

```
type(ratings_df)
```

```
     pyspark.sql.dataframe.DataFrame
```

```
ratings_df.count()
```

```
     981756
```

```
ratings_df.printSchema()
```

```
     root
      |-- book_id: integer (nullable = true)
      |-- user_id: integer (nullable = true)
      |-- rating: integer (nullable = true)
```

```
ratings_df.first()
```

```
     Row(book_id=1, user_id=314, rating=5)
```

```
ratings_df.show(5)
```

```
     +-------+-------+------+
     |book_id|user_id|rating|
     +-------+-------+------+
     |      1|    314|     5|
     |      1|    439|     3|
     |      1|    588|     5|
     |      1|   1169|     4|
     |      1|   1185|     4|
     +-------+-------+------+
     only showing top 5 rows
```

```
ratings_df.head(5)
```

```
     [Row(book_id=1, user_id=314, rating=5),
      Row(book_id=1, user_id=439, rating=3),
      Row(book_id=1, user_id=588, rating=5),
      Row(book_id=1, user_id=1169, rating=4),
      Row(book_id=1, user_id=1185, rating=4)]
```

```
ratings_df.select("book_id","rating").show(5)
```

```
     +-------+------+
     |book_id|rating|
     +-------+------+
     |      1|     5|
     |      1|     3|
     |      1|     5|
     |      1|     4|
     |      1|     4|
     +-------+------+
     only showing top 5 rows
```

```
ratings_df.filter("rating <= 3").show(5)
```

```
     +-------+-------+------+
     |book_id|user_id|rating|
     +-------+-------+------+
     |      1|    439|     3|
     |      1|   5461|     3|
     |      1|   7563|     3|
     |      1|   9246|     1|
     |      1|  20076|     3|
     +-------+-------+------+
     only showing top 5 rows
```

```
ratings_df.select("book_id","rating").filter("rating <= 3").show(5)
```

```
+-------+------+
|book_id|rating|
+-------+------+
|      1|     3|
|      1|     3|
|      1|     3|
|      1|     1|
|      1|     3|
+-------+------+
only showing top 5 rows
```

```
ratings_df.count()
```

```
981756
```

```
print(f"Total number of Ratings Records : {ratings_df.count()}")
```

```
Total number of Ratings Records : 981756
```

```
unique_user_count = ratings_df.select("user_id").distinct().count()
unique_user_count
```

```
53424
```

```
book_rating_less_or_three_count = ratings_df.filter("rating <= 3").count()
book_rating_less_or_three_count
```

```
331429
```

```
ratings_df.describe('book_id').show()
```

```
+-------+-----------------+
|summary|          book_id|
+-------+-----------------+
|  count|           981756|
|   mean|4943.275635697668|
| stddev|2873.207414896114|
|    min|                1|
|    max|            10000|
+-------+-----------------+
```

```
ratings_df.describe('book_id','rating').show()
```

```
+-------+-----------------+------------------+
|summary|          book_id|            rating|
+-------+-----------------+------------------+
|  count|           981756|            981756|
|   mean|4943.275635697668|3.8565335989797873|
| stddev|2873.207414896114|0.9839408559620033|
|    min|                1|                 1|
|    max|            10000|                 5|
+--------+----------------+------------------+
```

```
ratings_df.count()
```

```
981756
```

```
aaa = ratings_df.dropDuplicates()
```

```
aaa.count()
```

```
980112
```

```
rating_without_null = ratings_df.dropna().count()
```

```
rating_without_null
```

```
981756
```

```
ratings_df.dropna('any').count() # drop a row if it contains any nulls
```

```
981756
```

```
ratings_df.dropna('all').count() # drop a row if it contains any nulls
```

```
981756
```

```
ratings_df.agg({'rating':'max'}).show()
```

```
+-----------+
|max(rating)|
+-----------+
|          5|
+-----------+
```

```
ratings_df.groupby("rating").count().toPandas()
```

	rating	count
0	1	19575
1	3	248623
2	5	292961
3	4	357366
4	2	63231

```
ratings_df.groupby("rating").count().show()
```

```
+------+------+
|rating| count|
+------+------+
|     1| 19575|
|     3|248623|
|     5|292961|
|     4|357366|
|     2| 63231|
+------+------+
```

```
ratings_df.join(books_df, books_df.book_id == ratings_df.book_id).select("user_id","title"
```

```
+-------+--------------------+
|user_id|               title|
+-------+--------------------+
|    314|Harry Potter and ...|
|    439|Harry Potter and ...|
|    588|Harry Potter and ...|
|   1169|Harry Potter and ...|
|   1185|Harry Potter and ...|
+-------+--------------------+
only showing top 5 rows
```

```
ratings_df.orderBy("rating").show(5)
```

```
+-------+-------+------+
|book_id|user_id|rating|
+-------+-------+------+
|   6628|  39907|     1|
|   6631|  24498|     1|
|   6629|  47480|     1|
|   6627|  52717|     1|
|   6630|  27769|     1|
+-------+-------+------+
only showing top 5 rows
```

```
ratings_df.orderBy(ratings_df.rating.desc()).show(5)
```

```
+-------+-------+------+
|book_id|user_id|rating|
+-------+-------+------+
|   6627|  39631|     5|
|   6627|  48625|     5|
|   6627|  40962|     5|
|   6627|  34663|     5|
|   6627|  41337|     5|
+-------+-------+------+
only showing top 5 rows
```

```
ratings_df.orderBy("rating","book_id").show(115)
```

```
+-------+-------+------+
|book_id|user_id|rating|
+-------+-------+------+
|      1|   9246|     1|
|      1|  51480|     1|
|      2|  48687|     1|
|      2|  17643|     1|
|      2|  13794|     1|
|      2|   6063|     1|
|      3|  48687|     1|
|      3|  10246|     1|
|      3|  32305|     1|
|      3|  21733|     1|
|      3|  37284|     1|
|      3|  10610|     1|
|      3|  15604|     1|
|      3|  16569|     1|
|      3|   9246|     1|
|      3|  16377|     1|
|      3|  33065|     1|
|      3|  29703|     1|
|      3|  13794|     1|
|      3|    588|     1|
|      3|  11854|     1|
|      3|   4536|     1|
|      3|  52036|     1|
|      3|  23576|     1|
|      3|  10751|     1|
|      3|  25214|     1|
|      3|  10509|     1|
|      3|  10944|     1|
|      3|  49298|     1|
|      4|   4606|     1|
|      5|   3022|     1|
|      6|  18179|     1|
|      6|  18031|     1|
|      7|  13282|     1|
|      7|  51480|     1|
|      7|  23576|     1|
|      7|  12455|     1|
|      7|  11868|     1|
|      8|   3022|     1|
|      8|   3922|     1|
|      8|  18313|     1|
|      8|  17643|     1|
|      9|  11408|     1|
|      9|  38080|     1|
|      9|  11854|     1|
|      9|   7563|     1|
|      9|    951|     1|
|      9|  12466|     1|
|      9|  51480|     1|
|     10|   6063|     1|
|     10|  17643|     1|
|     11|  28824|     1|
|     12|  52036|     1|
|     12|  52007|     1|
|     14|  30681|     1|
ratings_df.withColumn("rating", ratings_df.rating*10).show(5)
```

```
+-------+-------+------+
|book_id|user_id|rating|
+-------+-------+------+
|      1|    314|    50|
|      1|    439|    30|
|      1|    588|    50|
|      1|   1169|    40|
|      1|   1185|    40|
+-------+-------+------+
only showing top 5 rows
```

```
new_dataset = ratings_df.withColumn("rating_ten", ratings_df.rating*10)
new_dataset.show(5)
```

```
+-------+-------+------+----------+
|book_id|user_id|rating|rating_ten|
+-------+-------+------+----------+
|      1|    314|     5|        50|
|      1|    439|     3|        30|
|      1|    588|     5|        50|
|      1|   1169|     4|        40|
|      1|   1185|     4|        40|
+-------+-------+------+----------+
only showing top 5 rows
```

```
ratings_df.show(5)
```

```
+-------+-------+------+
|book_id|user_id|rating|
+-------+-------+------+
|      1|    314|     5|
|      1|    439|     3|
|      1|    588|     5|
|      1|   1169|     4|
|      1|   1185|     4|
+-------+-------+------+
only showing top 5 rows
```

```
ratings_df.drop('rating').show(5)
```

```
+-------+-------+
|book_id|user_id|
+-------+-------+
|      1|    314|
|      1|    439|
|      1|    588|
|      1|   1169|
|      1|   1185|
+-------+-------+
only showing top 5 rows
```

```
ratings_df.show()
```

```
+-------+-------+------+
|book_id|user_id|rating|
+-------+-------+------+
|      1|    314|     5|
|      1|    439|     3|
|      1|    588|     5|
|      1|   1169|     4|
|      1|   1185|     4|
|      1|   2077|     4|
|      1|   2487|     4|
|      1|   2900|     5|
|      1|   3662|     4|
|      1|   3922|     5|
|      1|   5379|     5|
|      1|   5461|     3|
|      1|   5885|     5|
|      1|   6630|     5|
|      1|   7563|     3|
|      1|   9246|     1|
|      1|  10140|     4|
|      1|  10146|     5|
|      1|  10246|     4|
|      1|  10335|     4|
+-------+-------+------+
only showing top 20 rows
```

✓ 0s completed at 5:59 PM ● ×