In [1]:

```
import findspark
findspark.init()
```

In [2]:

```
import pyspark
from pyspark.sql import SparkSession
from pyspark.sql.types import *
from pyspark.sql.functions import *

# Configure spark session
spark = SparkSession\
    .builder\
    .master('local[2]')\
    .appName('AMAZON_BOOK')\
    .config('spark.jars.packages', 'org.mongodb.spark:mongo-spark-connector_2.12:2.4.1
    .config("spark.driver.memory", "5g")\
    .getOrCreate()
```

In [3]:

```
# MongoDB connection URI
mongo_uri = "mongodb://localhost:27017/AMAZON_BOOK.RATE"
# Read data from MongoDB collection into a DataFrame
df_rate = spark.read.format("mongo").option("uri", mongo_uri).load()
# Show the DataFrame
df_rate.show()
```

```
+----------+-----+--------------------+--------------+--------------------+-----+----
----------------+--------------------+
|        Id|Price|               Title|       User_id|                 _id|score|    
summary|                text|
+----------+-----+--------------------+--------------+--------------------+-----+----
----------------+--------------------+
|1558746153| NULL|Chicken Soup for ...| AEKP4FJRWRGZT|{6570367e824b9730...|  5.0|    
Helpful|Shows you what ot...|
|1882931173| NULL|Its Only Art If I...| AVCGYZL8FQQTD|{6570367e824b9730...|  4.0|Nice
collection o...|This is only for ...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  5.0|"Thi
s book hit th...|This book was ver...|
|0826414346| NULL|Dr. Seuss: Americ...|A30TK6U7DNS82R|{6570367e824b9730...|  5.0|   R
eally Enjoyed It|I don't care much...|
|0826414346| NULL|Dr. Seuss: Americ...|A3UH4UZ4RSVO82|{6570367e824b9730...|  5.0|Esse
ntial for eve...|"If people become...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  4.0| oh
this book was ok|well me and my fr...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  4.0| oh
this book was ok|well me and my fr...|
|0826414346| NULL|Dr. Seuss: Americ...|A2MVUWT453QH61|{6570367e824b9730...|  4.0|Phli
p Nel gives s...|Theodore Seuss Ge...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  5.0|Chic
ken Soup for ...|Chicken Soup for ...|
|0826414346| NULL|Dr. Seuss: Americ...|A22X4XUPKF66MR|{6570367e824b9730...|  4.0|Good
academic ove...|"Philip Nel - Dr....|
|1558746153| NULL|Chicken Soup for ...|A1M7N5V6W0Z0H7|{6570367e824b9730...|  5.0|More
stories abou...|Another great boo...|
|0826414346| NULL|Dr. Seuss: Americ...|A2F6NONFUDB6UK|{6570367e824b9730...|  4.0|One
of America's ...|"""Dr. Seuss: Ame...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  5.0|    
Heart-Warmer|Chicken Soup is a...|
|0826414346| NULL|Dr. Seuss: Americ...|A14OJS0VWMOSWO|{6570367e824b9730...|  5.0|A me
morably excel...|Theodor Seuss Gie...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  5.0|    
RACHEL'S REVIEW|One day my friend...|
|0826414346| NULL|Dr. Seuss: Americ...|A2RSSXTDZDUSH4|{6570367e824b9730...|  5.0|Acad
emia At It's ...|"When I recieved ...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  5.0|This
is one of th...|Chicken Soup for ...|
|0826414346| NULL|Dr. Seuss: Americ...|A25MD5I2GUIW6W|{6570367e824b9730...|  5.0|And
to think that...|"Trams (or any pu...|
|0826414346| NULL|Dr. Seuss: Americ...|A3VA4XFS5WNJO3|{6570367e824b9730...|  4.0|Fasc
inating accou...|As far as I am aw...|
|1558746153| NULL|Chicken Soup for ...|A1UMNA2JK9NELD|{6570367e824b9730...|  5.0|Klee
nex Needed to...|Warning: Do not r...|
+----------+-----+--------------------+--------------+--------------------+-----+----
----------------+--------------------+
only showing top 20 rows
```

In [4]:

```
df_rate = df_rate.drop('summary')
df_rate
```

Out[4]:

```
DataFrame[Id: string, Price: string, Title: string, User_id: string, _id: struct<oid:
string>, score: string, text: string]
```

In [5]:

```
df_rate.show()
```

```
+----------+-----+--------------------+--------------+--------------------+-----+----
----------------+
|        Id|Price|               Title|       User_id|                 _id|score|
text|
+----------+-----+--------------------+--------------+--------------------+-----+----
----------------+
|1558746153| NULL|Chicken Soup for ...| AEKP4FJRWRGZT|{6570367e824b9730...|  5.0|Show
s you what ot...|
|1882931173| NULL|Its Only Art If I...| AVCGYZL8FQQTD|{6570367e824b9730...|  4.0|This
is only for ...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  5.0|This
book was ver...|
|0826414346| NULL|Dr. Seuss: Americ...|A30TK6U7DNS82R|{6570367e824b9730...|  5.0|I do
n't care much...|
|0826414346| NULL|Dr. Seuss: Americ...|A3UH4UZ4RSVO82|{6570367e824b9730...|  5.0|"If
people become...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  4.0|well
me and my fr...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  4.0|well
me and my fr...|
|0826414346| NULL|Dr. Seuss: Americ...|A2MVUWT453QH61|{6570367e824b9730...|  4.0|Theo
dore Seuss Ge...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  5.0|Chic
ken Soup for ...|
|0826414346| NULL|Dr. Seuss: Americ...|A22X4XUPKF66MR|{6570367e824b9730...|  4.0|"Phi
lip Nel - Dr....|
|1558746153| NULL|Chicken Soup for ...|A1M7N5V6W0Z0H7|{6570367e824b9730...|  5.0|Anot
her great boo...|
|0826414346| NULL|Dr. Seuss: Americ...|A2F6NONFUDB6UK|{6570367e824b9730...|  4.0|"""D
r. Seuss: Ame...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  5.0|Chic
ken Soup is a...|
|0826414346| NULL|Dr. Seuss: Americ...|A14OJS0VWMOSWO|{6570367e824b9730...|  5.0|Theo
dor Seuss Gie...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  5.0|One
day my friend...|
|0826414346| NULL|Dr. Seuss: Americ...|A2RSSXTDZDUSH4|{6570367e824b9730...|  5.0|"Whe
n I recieved ...|
|1558746153| NULL|Chicken Soup for ...|          NULL|{6570367e824b9730...|  5.0|Chic
ken Soup for ...|
|0826414346| NULL|Dr. Seuss: Americ...|A25MD5I2GUIW6W|{6570367e824b9730...|  5.0|"Tra
ms (or any pu...|
|0826414346| NULL|Dr. Seuss: Americ...|A3VA4XFS5WNJO3|{6570367e824b9730...|  4.0|As f
ar as I am aw...|
|1558746153| NULL|Chicken Soup for ...|A1UMNA2JK9NELD|{6570367e824b9730...|  5.0|Warn
ing: Do not r...|
+----------+-----+--------------------+--------------+--------------------+-----+----
----------------+
only showing top 20 rows
```

In [6]:

```
df_rate = df_rate.drop('_id')
df_rate = df_rate.drop('Price')
df_rate
```

Out[6]:

```
DataFrame[Id: string, Title: string, User_id: string, score: string, text: string]
```

In [7]:

```
# Convert the "score" column to float
df_rate = df_rate.withColumn("score", col("score").cast("float"))
```

```
df_rate
```

Out[7]:

```
DataFrame[Id: string, Title: string, User_id: string, score: float, text: string]
```

In [8]:

```
from pyspark.sql.functions import isnan, when, count, col

# Check for null or NaN values in the "score" column
df_rate.select([count(when(isnan('score') | col('score').isNull(), 'score'))]).show()
```

```
+------------------------------------------------------------------+
|count(CASE WHEN (isnan(score) OR (score IS NULL)) THEN score END)|
+------------------------------------------------------------------+
|                                                             17922|
+------------------------------------------------------------------+
```

In [9]:

```
# Supprimez les lignes avec des valeurs NULL ou NaN dans la colonne "score"
df_rate = df_rate.na.drop(subset=["score"])
# Check for null or NaN values in the "score" column
df_rate.select([count(when(isnan('score') | col('score').isNull(), 'score'))]).show()
```

```
+------------------------------------------------------------------+
|count(CASE WHEN (isnan(score) OR (score IS NULL)) THEN score END)|
+------------------------------------------------------------------+
|                                                                 0|
+------------------------------------------------------------------+
```

In [10]:

```
df_rate = df_rate.drop('Text')
df_rate
```

Out[10]:

```
DataFrame[Id: string, Title: string, User_id: string, score: float]
```

In [11]:

```
!pip install scikit-surprise
```

```
Requirement already satisfied: scikit-surprise in c:\users\pc\anaconda3\lib\site-pack
ages (1.1.3)
Requirement already satisfied: joblib>=1.0.0 in c:\users\pc\anaconda3\lib\site-packag
es (from scikit-surprise) (1.2.0)
Requirement already satisfied: numpy>=1.17.3 in c:\users\pc\anaconda3\lib\site-packag
es (from scikit-surprise) (1.24.3)
Requirement already satisfied: scipy>=1.3.2 in c:\users\pc\anaconda3\lib\site-package
s (from scikit-surprise) (1.10.1)
```

In [12]:

```
!pip install scikit-learn
```

```
Requirement already satisfied: scikit-learn in c:\users\pc\appdata\roaming\python\pyt
hon311\site-packages (1.3.1)
Requirement already satisfied: numpy<2.0,>=1.17.3 in c:\users\pc\anaconda3\lib\site-p
ackages (from scikit-learn) (1.24.3)
Requirement already satisfied: scipy>=1.5.0 in c:\users\pc\anaconda3\lib\site-package
s (from scikit-learn) (1.10.1)
Requirement already satisfied: joblib>=1.1.1 in c:\users\pc\anaconda3\lib\site-packag
es (from scikit-learn) (1.2.0)
Requirement already satisfied: threadpoolctl>=2.0.0 in c:\users\pc\anaconda3\lib\site
-packages (from scikit-learn) (2.2.0)
```

In [13]:

```
# Sélectionner 22,000 lignes aléatoires
sample_df = df_rate.limit(22000)
```

```
# Convertir le DataFrame PySpark en Pandas DataFrame
pandas_df = sample_df.toPandas()
```

In [14]:

```
from surprise import Dataset, Reader

reader = Reader(rating_scale=(1, 5))
data = Dataset.load_from_df(pandas_df[['User_id', 'Title', 'score']], reader)
```

In [15]:

```
from surprise.model_selection import train_test_split

trainset, testset = train_test_split(data, test_size=0.2)
```

In [16]:

```
from surprise import KNNWithMeans

model = KNNWithMeans()
model.fit(trainset)
```

```
Computing the msd similarity matrix...
Done computing similarity matrix.
```

Out[16]:

```
<surprise.prediction_algorithms.knns.KNNWithMeans at 0x23f5ff783d0>
```

In [18]:

```
predictions = model.test(testset)
from surprise import accuracy
# Calcul de la RMSE (Root Mean Squared Error)
accuracy.rmse(predictions)
```

```
RMSE: 1.1866
```

Out[18]:

```
1.1865654453529382
```

In [ ]:

```

```