Assignment

November 7, 2024

1 Spark Preparation

We check if we are in Google Colab. If this is the case, install all necessary packages.

To run spark in Colab, we need to first install all the dependencies in Colab environment i.e. Apache Spark 3.3.2 with hadoop 3.3, Java 8 and Findspark to locate the spark in the system. The tools installation can be carried out inside the Jupyter Notebook of the Colab. Learn more from A Must-Read Guide on How to Work with PySpark on Google Colab for Data Scientists!

[9]:

```
try:
    import google.colab

    IN_COLAB = True
except:
    IN_COLAB = False
```

[10]:

```
if IN_COLAB:
    !apt-get install openjdk-8-jdk-headless -qq > /dev/null
    !wget -q https://dlcdn.apache.org/spark/spark-3.3.2/spark-3.3.2-bin-hadoop3.
 ↪tgz
    !tar xf spark-3.3.2-bin-hadoop3.tgz
    !mv spark-3.3.2-bin-hadoop3 spark
    !pip install -q findspark
    import os

    os.environ["JAVA_HOME"] = "/usr/lib/jvm/java-8-openjdk-amd64"
    os.environ["SPARK_HOME"] = "/content/spark"
```

2 Start a Local Cluster

[11]:

```
from pyspark.sql import SparkSession

# Initialize a Spark session with local mode
spark = SparkSession.builder.appName("LocalClusterApp").master("local[*]").
 ↪getOrCreate()
```

3 Spark Assignment

Based on the movie review dataset in 'netflix-rotten-tomatoes-metacritic-imdb.csv', answer the below questions.

Note: do not clean or remove missing data

[12]:
```
# Create a sample DataFrame
df = spark.read.csv("netflix-rotten-tomatoes-metacritic-imdb.csv", header=True)
df.show(5)
```

```
+--------------------+--------------------+--------------------+----------------
+---------------+----------------+--------------------+------------+------------
---+--------------------+--------------------+-----------+----------+-----------
----------+----------------+---------------+--------------------+----------+----
--------+--------------------+--------------------+--------------------+--------
------------+--------------------+----------+--------------------+--------------
------+--------------------+------------+
|               Title|               Genre|                Tags|
Languages|Series or Movie|Hidden Gem Score|Country Availability|     Runtime|
Director|              Writer|              Actors|View Rating|IMDb Score|Rotten
Tomatoes Score|Metacritic Score|Awards Received|Awards Nominated For|
Boxoffice|Release Date|Netflix Release Date|    Production House|        Netflix
Link|           IMDb Link|             Summary|IMDb Votes|               Image|
Poster|        TMDb Trailer|Trailer Site|
+--------------------+--------------------+--------------------+----------------
+---------------+----------------+--------------------+------------+------------
---+--------------------+--------------------+-----------+----------+-----------
----------+----------------+---------------+--------------------+----------+----
--------+--------------------+--------------------+--------------------+--------
------------+--------------------+----------+--------------------+--------------
------+--------------------+------------+
|   Lets Fight Ghost|Crime, Drama, Fan…|Comedy Programmes…|Swedish, Spanish|
Series|             4.3|            Thailand|< 30 minutes|Tomas Alfredson|John
Ajvide Lindq…|Kåre Hedebrant, P…|          R|       7.9|
98.0|            82.0|           74.0|                57.0|$2,122,065| 12 Dec
2008|          2021-03-04|Canal+, Sandrew
M…|https://www.netfl…|https://www.imdb…|A med student wit…|
205926.0|https://occ-0-470…|https://m.media-a…|                NULL|
NULL|
|HOW TO BUILD A GIRL|              Comedy|Dramas,Comedies,F…|         English|
Movie|             7.0|              Canada|    1-2 hour|  Coky Giedroyc|
Caitlin Moran|Paddy Considine, …|          R|       5.8|                79.0|
69.0|            1.0|                NULL|   $70,632| 08 May 2020|
2021-03-04|Film 4, Monumenta…|https://www.netfl…|https://www.imdb…|When
nerdy Johann…|
2838.0|https://occ-0-108…|https://m.media-a…|https://www.youtu…|
YouTube|
|         Centigrade|     Drama, Thriller|           Thrillers|         English|
Movie|             6.4|              Canada|    1-2 hour|  Brendan Walsh|Brendan
Walsh, Da…|Genesis Rodriguez…|    Unrated|       4.3|                NULL|
46.0|           NULL|                NULL|   $16,263| 28 Aug 2020|
2021-03-04|
NULL|https://www.netfl…|https://www.imdb…|Trapped in a froz…|
1720.0|https://occ-0-108…|https://m.media-a…|https://www.youtu…|
YouTube|
|              ANNE+|               Drama|TV Dramas,Romanti…|         Turkish|
Series|             7.7| Belgium,Netherlands|< 30 minutes|           NULL|
NULL|Vahide Perçin, Go…|       NULL|       6.5|                NULL|
NULL|            1.0|                NULL|      NULL| 01 Oct 2016|
2021-03-04|                NULL|https://www.netfl…|https://www.imdb…|Upon
moving into …|    1147.0|https://occ-0-148…|https://m.media-a…|
NULL|        NULL|
|              Moxie|Animation, Short,…|Social Issue Dram…|         English|
Movie|             8.1|Lithuania,Poland,…|    1-2 hour|  Stephen Irwin|
NULL|        Ragga Gudrun|       NULL|       6.3|                NULL|
NULL|           NULL|                 4.0|      NULL| 22 Sep 2011|
2021-03-04|
NULL|https://www.netfl…|https://www.imdb…|Inspired by her m…|
63.0|https://occ-0-403…|https://m.media-a…|               NULL|
NULL|
+--------------------+--------------------+--------------------+----------------
+----------------+----------------+--------------------+------------+------------
---+--------------------+--------------------+------------+----------+-----------
----------+----------------+----------------+--------------------+----------+----
--------+--------------------+--------------------+--------------------+--------
------------+--------------------+----------+--------------------+--------------
------+--------------------+------------+
only showing top 5 rows
```

3.1 What is the maximum and average of the overall hidden gem score?

```
[44]: df.select("Hidden Gem Score").summary("max", "mean").show()
```

```
+-------+-----------------+
|summary| Hidden Gem Score|
+-------+-----------------+
|    max|              9.8|
|   mean|5.937551386501234|
+-------+-----------------+
```

3.2 How many movies that are available in Korea?

[66]:
```
from pyspark.sql.functions import col

df.select("Country Availability").filter(col("Country Availability").
 ↪contains("Korea")).count()
```

[66]:
```
4845
```

3.3 Which director has the highest average hidden gem score?

[15]:
```
from pyspark.sql.functions import avg

df.groupBy("Director").agg(avg("Hidden Gem Score")).sort("avg(Hidden Gem␣
 ↪Score)", ascending=False).show(1)
```

```
+-----------+---------------------+
|   Director|avg(Hidden Gem Score)|
+-----------+---------------------+
|Dorin Marcu|                  9.8|
+-----------+---------------------+
only showing top 1 row
```

3.4 How many genres are there in the dataset?

[37]:
```
from pyspark.sql.functions import split, explode

df_genres = df.withColumn("Genre", split(df["Genre"], ", "))
df_genres_exploded = df_genres.select(explode("Genre").alias("Genre"))
df_genres_exploded.select("Genre").distinct().count()
```

[37]:
```
28
```