Học phần Quản trị Dữ liệu lớn: Bài thực hành số 5 - Machine learning với Spark

Phạm Tiến Lâm, Đặng Văn Báu

- Khai báo thư viện

```
from pyspark.sql import SparkSession
from pyspark.ml.recommendation import ALS
from pyspark.sql import Row
from pyspark.sql.functions import lit
```

- Load movieID và movieName từ file u.item thành một Dict

```
1  :  Toy Story (1995)
2  :  GoldenEye (1995)
3  :  Four Rooms (1995)
4  :  Get Shorty (1995)
5  :  Copycat (1995)
6  :  Shanghai Triad (Yao a yao yao dao waipo qiao) (1995)
7  :  Twelve Monkeys (1995)
8  :  Babe (1995)
9  :  Dead Man Walking (1995)
10  :  Richard III (1995)
```

- Sử dụng SparkSession

```
# Create a SparkSession
spark = SparkSession.builder.appName("MovieRecs").getOrCreate()

# This line is necessary on HDP 2.6.5:
spark.conf.set("spark.sql.crossJoin.enabled", "true")
```

- Khởi tạo hàm ParseInput từ file u.data và convert dữ liệu thành các rows và convert thành một RDD.

```
# Get the raw data
lines = spark.read.text("ml-100k/u.data").rdd

# Convert it to a RDD of Row objects with (userID, movieID, rating)
ratingsRDD = lines.map(parseInput)
```

- Chuyển đổi RDD thành dataframe và cache nó.

```
ratings = spark.createDataFrame(ratingsRDD).cache()
```

- Sử dụng ALS trong Spark để train model

```
# Create an ALS collaborative filtering model from the complete data set
als = ALS(maxIter=5, regParam=0.01, userCol="userID", itemCol="movieID", ratingCol="rating")
model = als.fit(ratings)
```

- Hiển thị các phim mà userID = 0 đã rate

```
# Print out ratings of user 0:
print("\nRatings for user ID 0:")
userRatings = ratings.filter("userID = 0")
for rating in userRatings.collect():
    print(movieNames[rating['movieID']], rating['rating'])
```

```
Ratings for user ID 0:
Star Wars (1977) 5.0
Empire Strikes Back, The (1980) 5.0
Gone with the Wind (1939) 1.0
```

- In ra tất cả các phim đã được rate trên 100 lần

```
# Find movies rated more than 100 times
ratingCounts = ratings.groupBy("movieID").count().filter("count > 100")
```

```
ratingCounts.take(10)
```

```
[Row(movieID=474, count=194),
 Row(movieID=29, count=114),
 Row(movieID=65, count=115),
 Row(movieID=191, count=276),
 Row(movieID=418, count=129),
 Row(movieID=222, count=365),
 Row(movieID=293, count=147),
 Row(movieID=270, count=136),
 Row(movieID=367, count=170),
 Row(movieID=705, count=137)]
```

- Tạo một dataframe có tên là popularMovies từ ratingCounts theo cột movieID. Và khởi tạo thêm 1 cột có tên là userID có giá trị bằng 0

```
popularMovies = ratingCounts.select("movieID").withColumn('userID', lit(0))
```

```
1 popularMovies.take(10)
```

```
[Row(movieID=474, userID=0),
 Row(movieID=29, userID=0),
 Row(movieID=65, userID=0),
 Row(movieID=191, userID=0),
 Row(movieID=418, userID=0),
 Row(movieID=222, userID=0),
 Row(movieID=293, userID=0),
 Row(movieID=270, userID=0),
 Row(movieID=367, userID=0),
 Row(movieID=705, userID=0)]
```

- Biến đổi popularMovies theo model đã train ở bước trên thành spark.sql.dataframe

```
recommendations = model.transform(popularMovies)

type(recommendations)
```

```
1 recommendations.take(10)
```

```
[Row(movieID=148, userID=0, prediction=2.6911063194274902),
 Row(movieID=471, userID=0, prediction=3.342437744140625),
 Row(movieID=496, userID=0, prediction=2.594545364379883),
 Row(movieID=243, userID=0, prediction=3.0868818759918213),
 Row(movieID=31, userID=0, prediction=4.521707534790039),
 Row(movieID=137, userID=0, prediction=2.0551247596740723),
 Row(movieID=451, userID=0, prediction=2.0275745391845703),
 Row(movieID=65, userID=0, prediction=2.2380459308624268),
 Row(movieID=879, userID=0, prediction=3.3363752365112305),
 Row(movieID=53, userID=0, prediction=2.753848075866699)]
```

- Sắp xếp dữ liệu theo thứ tự giảm dần trong recommendations theo cột ‘prediction’

```
Terminator, The (1984) 5.857853889465332
Star Trek: The Wrath of Khan (1982) 5.844795227050781
Nightmare on Elm Street, A (1984) 5.638725757598877
Aliens (1986) 5.459494590759277
Chasing Amy (1997) 5.332832336425781
Die Hard (1988) 5.232217788696289
Alien (1979) 5.196457386016846
Tombstone (1993) 5.188972473144531
Terminator 2: Judgment Day (1991) 5.17092752456665
Game, The (1997) 5.1444010734558105
Raiders of the Lost Ark (1981) 5.055807590484619
Star Trek IV: The Voyage Home (1986) 5.047184944152832
Batman (1989) 5.025223731994629
Star Trek: First Contact (1996) 4.995238304138184
Star Wars (1977) 4.982772350311279
Empire Strikes Back, The (1980) 4.9452314376831055
Return of the Jedi (1983) 4.925889015197754
Die Hard: With a Vengeance (1995) 4.911350250244141
Searching for Bobby Fischer (1993) 4.89263391494751
Sneakers (1992) 4.888552665710449
```