databricksHW_01

DSCI 417 – Homework 01

Sean Graham

```
import math
from pyspark.sql import SparkSession
from pyspark.mllib.random import RandomRDDs


spark = SparkSession.builder.getOrCreate()
sc = spark.sparkContext
```

Problem 1: Terminology

1. Scala
2. SparkSession
3. SparkContext
4. Resilient Distributed Dataset
5. Partitions
6. Transformation
7. Action
8. Transformation
9. Action
0. Transformation
1. Action
2. List
3. Master
4. Workers
5. Driver
6. Executor

Problem 2: Working with a Numerical RDD

```
random_rdd = RandomRDDs.uniformRDD(sc, size=1200000, seed=1)

print('Sum:     ', random_rdd.sum())
print('Mean:    ', random_rdd.mean())
print('Std Dev: ', random_rdd.stdev())
print('Minimum: ', random_rdd.min())
print('Maximum: ', random_rdd.max())
```

```
Sum:      599075.0656203285
Mean:     0.49922922135026476
Std Dev:  0.28869756091861193
Minimum:  1.0351479373671424e-07
Maximum:  0.9999991929309536
```

```
print('Number of Partitions: ', random_rdd.getNumPartitions())
print('Size of Partitions:   ')
print(random_rdd.glom().map(len).collect())
```

```
Number of Partitions:  4
Size of Partitions:
[300000, 300000, 300000, 300000]
```

Problem 3: Transformations

```
scaled_rdd = random_rdd.map(lambda x : x*10)

print('Sum:     ', scaled_rdd.sum())
print('Mean:    ', scaled_rdd.mean())
print('Std Dev: ', scaled_rdd.stdev())
print('Minimum: ', scaled_rdd.min())
print('Maximum: ', scaled_rdd.max())
```

```
Sum:      5990750.656203327
Mean:     4.99229221350273
Std Dev:  2.886975609186112
Minimum:  1.0351479373671424e-06
Maximum:  9.999991929309536
```

```
log_rdd = scaled_rdd.map(lambda x : math.log(x))

print('Sum:     ', log_rdd.sum())
print('Mean:    ', log_rdd.mean())
print('Std Dev: ', log_rdd.stdev())
print('Minimum: ', log_rdd.min())
print('Maximum: ', log_rdd.max())
```

```
Sum:      1559584.3948350847
Mean:     1.2996536623625732
Std Dev:  1.0035829379691585
Minimum:  -13.780966206806882
Maximum:  2.3025842859246737
```

Problem 4: Calculating SSE

```
pairs_raw = sc.textFile('/FileStore/tables/pairs_data.txt')

print(pairs_raw.count())
```

```
12743548
```

```
for row in pairs_raw.take(5):
    print(row)
```

```
12.3 12.1
```

```
9.1 8.7
9.3 9.9
8.5 8.5
11.2 10.8
```

```
def process_line(row):
    items = row.split(' ')
    return [float(items[0]), float(items[1])]

pairs = pairs_raw.map(process_line)

for row in pairs.take(5):
    print(row)
```

```
[12.3, 12.1]
[9.1, 8.7]
[9.3, 9.9]
[8.5, 8.5]
[11.2, 10.8]
```

```
square_rdd = pairs.map(lambda x : (x[1] - x[0])**2)
SSE = square_rdd.sum()

print(SSE)
```

```
4597380.190042952
```

Problem 5: Calculating r-Squared

```
mean_rdd = pairs.map(lambda x : x[0])
mean = mean_rdd.mean()

print(mean)
```

```
10.00013136059118
```

```
difference_rdd = pairs.map(lambda x : (x[0] - mean)**2)
SST = difference_rdd.sum()

print(SST)
```

```
24980514.859974924
```

```
r2 = 1 - SSE / SST

print(r2)
```

```
0.815961351644953
```

Problem 6: NASA Server Logs

```
nasa = sc.textFile('/FileStore/tables/NASA_server_logs_Aug_1995.txt')

print(nasa.count())
```

```
1569888
```

```
for row in nasa.take(5):
    print(row)
```

```
in24.inetnebr.com [01/Aug/1995:00:00:01] "GET /shuttle/missions/sts-68/news/sts-68-mcc-05.txt" 200 1839
```

```
uplherc.upl.com [01/Aug/1995:00:00:07] "GET /" 304 0
uplherc.upl.com [01/Aug/1995:00:00:08] "GET /images/ksclogo-medium.gif" 304 0
uplherc.upl.com [01/Aug/1995:00:00:08] "GET /images/MOSAIC-logosmall.gif" 304 0
uplherc.upl.com [01/Aug/1995:00:00:08] "GET /images/USA-logosmall.gif" 304 0

Number of GET requests:   1565812
Number of POST requests:  111
Number of HEAD requests:  3965
```