ex392

August 12, 2022

[ ]:
```
from pyspark import SparkConf, SparkContext
conf = SparkConf().setAppName("ex392")
sc = SparkContext(conf=conf)
```

[2]:
```
inputPath = "data/Ex39bis/data/"
outputPath = "out392/"
```

[3]:
```
inputRDD = sc.textFile(inputPath)
```

[4]:
```
datesRDD = inputRDD\
    .filter(lambda line : float(line.split(",")[2])>50.0)\
        .map(lambda line : (line.split(",")[0],line.split(",")[1]))\
            .groupByKey()\
                .mapValues(lambda dates: list(dates))
##stesso procedimento della versione precedente
```

[5]:
```
#colleziono tutti gli ID
sensorsRDD = inputRDD\
    .map(lambda line : line.split(",")[0])
#rimuovo gli ID buoni e rimango solo con quelli che non hanno superato la␣
 ↪treshold, e li mappo con una lista vuota in un pair RDD
badSensorsRDD = sensorsRDD.subtract(datesRDD.keys())\
    .map(lambda sensorid : (sensorid, list()))
```

[6]:
```
finalRDD = badSensorsRDD.union(datesRDD) #infine faccio l'unione dei due RDD
```

[7]:
```
finalRDD.saveAsTextFile(outputPath)
```