ex42

August 12, 2022

[ ]:
```
from pyspark import SparkContext, SparkConf
conf = SparkConf().setAppName("ex42")
sc = SparkContext(conf=conf)
```

[23]:
```
inputPathQuestions = "data/Ex42/data/questions.txt"
inputPathAnswers = "data/Ex42/data/answers.txt"
outputPath = "out42/"
```

[24]:
```
questionsRDD = sc.textFile(inputPathQuestions)
answersRDD = sc.textFile(inputPathAnswers)
```

[25]:
```
questionPairRDD = questionsRDD.map(lambda line : (line.split(",")[0], line.
 ↪split(",")[2]))
answerPairRDD = answersRDD.map(lambda line : (line.split(",")[1], line.
 ↪split(",")[3]))
```

[26]:
```
questionAnswersRDD = questionPairRDD.cogroup(answerPairRDD)
```

[27]:
```
finalRDD = questionAnswersRDD.mapValues(lambda value : (list(value[0]),␣
 ↪list(value[1])))
```

[28]:
```
finalRDD.saveAsTextFile(outputPath)
```