Name: ___________________________ USC ID: ______________________

Quiz 13: Apache Spark (10 points. 10 minutes)

1. [8 points] Consider the following Spark code:

```
lines = sc.textFile(“hello.txt”)
lines1 = lines.filter(lambda x: “reduce” in x)
words = lines1.flatMap(lambda x: x.split(' '))
kvs = words.map(lambda x: (x, 1))
counts = kvs.reduceByKey(lambda x, y: x + y)
```

Suppose the “*hello.txt*” file has the following 3 lines:

```
map or reduce
map and only map
reduce after map
```

What are the contents of these RDDs: **lines, lines1, words, kvs, and counts**?

RDD	Content
lines	['map or reduce', 'map and only map', 'reduce after map']
lines1	['map or reduce', 'reduce after map']
words	['map', 'or', 'reduce', 'reduce', 'after', 'map']
kvs	[('map', 1), ('or', 1), ('reduce', 1), ('reduce', 1), ('after', 1), ('map', 1)]
counts	[('map', 2), ('after', 1), ('reduce', 2), ('or', 1)]

2. [2 points] Consider the following Spark code:

```
ds1 = sc.parallelize([(1,2), (2,3), (3,4)])
ds2 = sc.parallelize([(2,4), (3,6), (4,8)])
joinRes = ds1.join(ds2)
fullRes = ds1.fullOuterJoin(ds2)
```

What are the contents of **joinRes** and **fullRes**?

RDD	Content
joinRes	[(2, (3, 4)), (3, (4, 6))]
fullRes	[(1, (2, None)), (2, (3, 4)), (3, (4, 6)), (4, (None, 8))]

0.5 point for each entry