Name: _____ USC ID: _____

INF 551 – Fall 2017 (Morning section)

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

1. [4 points] Write a Sparky script that uses **only** aggregate() to implement mean() in Spark (example of using mean() is shown below).

```
data = sc.parallelize([1, 2, 3, 4, 5], 2)
m = data.mean()
```

output= data.aggregate((0,0), lambda U,v:(U[0] + v, U[1] + 1), lambda U,V:(U[0] + V[0], U[1] + V[1]))

print (float(output[0])/output[1])

2. [3 points] Write a Sparky script that uses **only** mapValues() and reduceByKey() to obtain the same result as countByKey() (example of which is shown below).

```
d = [('hello', 1), ('world', 1), ('hello', 2), ('this', 1), ('world',0)]
data = sc.parallelize(d, 2)
data.countByKey()
```

rdd2 = data.mapValues(lambda x: 1).reduceByKey(add).foreachPartition(printf)

3. [3 points] Write a Spark script to find all integers in a given RDD that is **divisible by 3**. The rdd is shown below.

```
rdd = sc.parallelize([3, 5, 7, 9, 12], 2)
```

result = rdd.filter(lambda x: x % 3 == 0).foreachPartition(printf)