Name: _____ USC ID: _____

INF 551 – Fall 2018 (Afternoon section)

Quiz 10: Spark (15 points), 15 minutes

Note: 5 points extra credit question!

1. [10 points] Write the Python Spark code for counting the number of occurrences of words in a text file "input.txt". **Requirements**: ignore words that contain no more than 2 characters.

```
from operator import add
lines = sc.textFile('input.txt')
result = lines.flatMap(lambda s: s.split()).filter(lambda x: len(x)>2)\
.map(lambda x: (x,1)).reduceByKey(add)
```

2. [5 points**, Extra Credit**!] Consider two RDDs: ds1 and ds2, each containing a list of key-value pairs. Write a Spark code that computes the same result as ds1.leftOuterJoin(ds2) by using other transformations (not join) seen in class.

```
data1 = ds1.map(lambda x, y: (x, ('ds1', y)))
data2 = ds2.map(lambda x, y: (x, ('ds2', y)))

Def leftOuterJoin(key, values):
    ds1_data = []
    ds2_data = []
    result = []
    for value in values:
        if value[0] == 'ds1':
            ds1_data.append(value[1])
        else:
            ds2_data.append(value[1])

    for ds1element in ds1_data:
        if len(ds2_data) == 0:
            result.append(key, (ds1element, None))
        else:
            for ds2element in ds2_data:
                result.append(key, (ds1element, ds2element))
    return result

ouput = data1.union(data2).groupByKey().flatMap(leftOuterJoin)
```