Name: _____ USC ID: _____

INF 551 – Fall 2018 (Morning 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 the following words: "and", "the", "a".

   Sample answer 1:

```
lines = sc.textFile("input.txt")
rdd = lines.flatMap(lambda s:s.split()).filter(lambda x:(x!="and" and x!="a" and
x!="the")).map(lambda x:(x,1)).countByKey()
```

   Sample answer 2:

```
def mapper(x):
    res=[]
    words=x.split()
    for each in words:
        if each!="and" and each!="a" and each!="the":
            res.append((each,1))
    return res
lines = sc.textFile("input.txt")
rdd = lines.flatMap(lambda x:mapper(x)).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.join(ds2) by using other transformations (i.e., not join) seen in class.

   Sample solution:

```
rdd1=ds1.map(lambda x:(x[0],("ds1",x[1])))
rdd2=ds2.map(lambda x:(x[0],("ds2",x[1])))
rdd3=rdd1.union(rdd2)
rdd4=rdd3.groupByKey()
def join_fun(x):
    result = []
    for tuple1 in x[1]:
        if tuple1[0]=="ds1":
            for tuple2 in x[1]:
                if tuple2[0]=="ds2":
                    result.append((x[0],(tuple1[1],tuple2[1])))
    return result
rdd5=rdd4.flatMap(lambda x:join_fun(x))
```