Name: _____ USC ID: _____

INF 551 – Fall 2017 (Morning section)

Quiz 12: Hadoop MapReduce (10 points), 15 minutes

Consider a sales data stored as a csv file (fields are category, state, year, and amount):

Cell,CA,2017,100
Cell,NY,2017,150
Laptop,CA,2017,200
Laptop,NY,2017,180
Cell,CA,2018,100

Consider computing the answer to the following SQL query using Hadoop MapReduce.

```
Select state, avg(amount)
From sales
Where year = 2017
Group by state
Having count(*) > 1
```

1. [5 points] Write the logic of **map** function in pseudocode. What will the map function output for the data above?

   For each k,v that we get inside the map function:
   Split v based on comma and we would get array of tokens
   If token[2] == "2017":
   output(token[1], int(token[3]))

   This will generate: [(CA, 100), (NY, 150), (CA, 200), (NY,180)]

2. [5 points] Write the logic of **reduce** function in pseudocode. What will the reduce function output for the data above?

   Count = 0
   Sum = 0

   For each k, list of values we get inside the reduce function

   Loop over all the values:
   Sum += value
   Count += 1

   If count > 1:
   Output(k, sum/count)

   This will generate [(CA, 150), (NY,165)]