Name: _____ USC ID: _____

Quiz 1: Hadoop MapReduce (10 points), 15 minutes

Consider an employee table stored as a comma-separated file, where each line is an employee record (id, name, age, and gender). The following shows an example file "Emp":

```
100,John,25,M
200,Mary,23,F
300,David,23,M
400,Bill,26,M
500,Jennifer,20,F
600,Maria,28,F
```

Consider using Hadoop MapReduce (without using combiner) to implement the following SQL query:

```
SELECT gender, count(*)
FROM Emp
WHERE age > 22
GROUP BY gender
```

1. [3 points] Implement the **map** function in pseudocode. Assume that input key to map function is line offset and value is line content (similar to WordCount program seen in class).

```
map (key, value):
        Extract (id, name, age, gender) by tokenizing from line content.  (1 point)
        If (age > 22)                                   (1 point)
                output (gender, 1)                      (1 point)
```

2. [2 points] State all key-value pairs output by the Map tasks for the example file "Emp".

(M, 1), (F,1), (M,1), (M, 1), (F, 1) (2 points, less or more than 5 outputs will have 1 point deduction)

3. [3 points] Implement the **reduce** function in pseudocode.

```
reduce(key, values):
    // key: gender; values: an iterator over count with same gender
    result = 0
    for each value v  in values:                (1 point)
        result += v;                            (1 point)

    output(key, result)                         (1 point)
```

4. [2 points] State the **input** and **output** for each call to the reduce function for the "Emp" file above.

Input : (M, [1,1,1]) (F, [1,1]) (2 points, 1 point for each output)
Output: (M, 3) (F, 2) (2 points, 1 point for each output)