Data Mining Quiz April 13, 2015

1. Explain what each of the following Python built-in functions do: `filter`, `map`, `reduce`.

2. What is a lambda expression? What is the syntax for lambda expressions in Python?

3. What will be the output of the following Python code? (Show steps of calculation)

```
a=[3,1,4,1,5,9]

def subt(x,y):
        return x-y

C = reduce(subt,x)

print C
```

4. What will be the output of the following Python code?

```
a=[1,2,3,4,5]
b=map(lambda a: a * a, a)
print b
c=filter(lambda x: (x > 12), b)
print c
C=reduce(lambda c,y: c+y,c)
print C
```

5. What is a *Resilient Distributed Dataset*?

6. What is a *Spark Context*? Also, what does the `parallelize()` method of a Spark Context do?

7. What is the difference between `map()` and `flatMap()`? In particular, what is the output of the following Python Spark script?

```
sc = SparkContext("local", "Simple App")

lines = sc.parallelize(["one one","two three four", "two three"])
```

```
words1 = lines.flatMap(lambda line: line.split(" "))
words2 = lines.map(lambda line: line.split(" "))

print words1.collect()
print words2.collect()
```

8. What does the `collect()` method of an RDD do?

9. What is *Gradient Descent*? Provide the update rule used in gradient descent.

10. Beginning with an initial guess of $x_{min} = 3$, how would you minimize the function $f(x) = x^2 + 1$ using gradient descent?