Name: _____ USC ID: _____

INF 551 – Fall 2017 (Afternoon section)

Quiz 13: Apache Spark (10 points), 10 minutes

1. [5 points] Write a Spark script that uses **only** aggregateByKey() to compute the group averages (where each group corresponds to a unique key in the input dataset) for the following RDD:

   rdd = sc.parallelize([(1,2), (1,3), (2,2), (1,4), (3,5), (2, 4), (1, 5), (2, 6)], 2)

   What is its output on the above dataset?

   rdd1 = rdd.aggregateByKey((0,0),lambda u, x: (u[0]+x, u[1]+1), lambda u, v: (u[0]+v[0],u[1]+v[1])).map(lambda (x,(y, z)): (x, float(y)/z))

   output: [(2, 4.0), (1, 3.5), (3, 5.0)]

2. [5 points] Write a Spark script that uses **only** mapValues() and flatMap() implement the following flatMapValues().

   rdd = sc.parallelize([(1, "hello world"), (2, "hello this world")])
   rdd2 = rdd.flatMapValues(lambda s: s.split())

   What is the content of rdd2?

   rdd2 = rdd.mapValues(lambda s: s.split()).flatMap(lambda (k, s): [(k, i) for i in s])

   content of rdd2:

   [(1, 'hello'), (1, 'world'), (2, 'hello'), (2, 'this'), (2, 'world')]