**Big Data Analytics – CS7070**

**Programming Project #2**

The goal of this project is to write a Spark program to compute the clustering coefficient of every node of a graph. The four different phases of the project, their expected subgoals, and their submission deadlines are as follows. A small graph, called TinyDataSet, is included at the end of this description. A somewhat larger test graph, called SmallDataSet, will be made available before the due date.

1. Phase-1: Write a program for Spark (in PySpark, Scala, or Spark-JAVA) that takes as input a list of edges of a graph, and outputs a node-list representation of the graph. In the output, for each node (=key) the value is a list of all the nodes to which the key node is connected.
   1. Items to be submitted: (i) your program source code, (ii) A list of input records used by your program, (iii) Output of your program for the TinyDataSet, (iv) Output of your program for the SmallDataSet.
   2. Due Date: March 23rd, 2020, 9PM.
2. Phase-2: Write a program for Spark (in PySpark, Scala, or Spark-JAVA) that takes as input a graph in the list representation (generated in phase-1 above) and produces a list of records such that the key is a node number and the value is the 2-hop projection of the node in the graph.
   1. Items to be submitted: (i) your program source code, (ii) Output of your program for the graph representation generated in Phase-2 for the TinyDataSet, (iii) Output of your program for the output generated by Phase-1 for the SmallDataSet.
   2. Due Date: March 26th, 2020, 9PM.
3. Phase-3: Write a program for Spark (in PySpark, Scala, or Spark-JAVA) that takes as input a graph in the form of 2-hop projection (value) for each node (key) and produces a list of (key, value) pairs such that for each node (key), the value is all the triangles incident on that node.
   1. Items to be submittwaed: (i) yourprogram source code, (ii) Output of your program for the graph of TinyDataSet, and (ii) Output of your program for the graph of SmallDataSet.
   2. Due Date: March 31st, 2020, 9PM.
4. Phase-4: Work out on a piece of paper how clustering-coefficient for a node of a graph may be computed from the output obtained after Phase-3. Compute the clustering coefficient for at least 5 nodes of the TinyDataSet and show all your work.
   1. Items to submit: Algorithm/methodology used by you to compute the clustering coefficient, (ii) all the work showing computation of clustering coefficient for five nodes.
   2. Due Date: April 3rd, 2020, 9PM.

TinyDataSet Graph: (List of edges)

7 10

7 8

7 4

8 9

8 5

9 5

9 10

10 6

4 5

5 6

4 6

1 4

1 3

2 3

2 6

3 4

3 6