University of Puerto Rico
Mayagüez Campus

Department of Computer Science and Engineering

CIIC 5995/8995: Big Data Analytics
Final Exam

Name: Stephan Elias Remy
Student ID: 802-12-2205

TOTAL: /100

Section: CIIC5995:116 **June 29, 2017**

Answer the following questions.

I) **(25 pts)<u>Problem I: MapReduce</u>**: Briefly explain the concept of MapReduce, and provide an example on what problems you can solve with it.

Map-Reduce is a data processing technique for processing data throughout a distributed system. It has two main phases: Map and Reduce. During the first phase, map, some mapping nodes take some data and converts it into a set of key-value pairs and pass the newly formed data to a set of reducing nodes. In the second phase, reduce, the reducing nodes take the data given to them and start to combine the data into one final output. An example of problem solved with Map-Reduce would be counting how many tweets have a certain keyword. The mappers would create a key-value pair in which the keyword would be the key and how many times it appeared in the tweet is the value. The reducers then would get all the values corresponding to the key of the keyword and then add all the values. At the end it would output another key-value pair which would be the keyword and the total times the word appeared.

II) **(25 pts)<u>Problem II: Spark RDD</u>**: Briefly explain the concept of the Spark RDDs and contrast their performance with that of MapReduce.

In Spark an RDD stands for “Resilient Distributed Dataset”. These RDDs are kept in memory and are read-only therefore any transformation does not modify the RDD instead it creates a new one. In addition, any transformation operation on the RDD is lazy. This means that even though the operation was called and used, Spark won’t evaluate anything until an action operation, collect as an example, is a called. Spark has a significant performance advantage over Map-Reduce for two main reasons. The data is kept in memory therefore no data is being transferred from node to node and the computation follows a directed acyclic graph model. This means that the action operations are pipelined along an operator pipeline.

III) **(25 pts)<u>Problem III: Supervised Learning</u>**: Briefly explain the concept of Supervised Learning, and provide an example on what problems you can solve with it.

Supervised Learning is a technique that allows algorithms to "learn" how to make certain predictions based on previous data. A set of training data is usually provided so the algorithm can infer upon the real data. Some applications of this would-be regression which helps predict a continuous value, classification which could be applied to verifying if a person is happy or not in a photograph, and more. A more formal example of what can be solved with Supervised learning techniques train the algorithm to classify if the tissue in an image is healthy or cancerous.

IV) **(25 pts)<u>Problem IV: Unsupervised Learning</u>**: Briefly explain the concept of Unsupervised Learning, and provide an example on what problems you can solve with it.

Unsupervised learning is similar to supervised learning in the sense that it is also a technique for making predictions on data. The difference between the two is that unsupervised learning techniques do not receive a labeled or structured data. Some examples of its application could be in outlier detection which detects if the input data is somehow very different that the rest and in cluster analysis which is commonly used to detect hidden patterns within the data.