University of Puerto Rico
Mayagüez Campus

Department of Computer Science and Engineering

CIIC 5995/8995: Big Data Analytics
Final Exam

Name: Jose Garcia Negron TOTAL: /100

Student ID: 802-12-2645

Section: _____ **June 29, 2017**

Answer the following questions.

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

MapReduce is a model for processing big amounts of data. It does it by applying a map operation to all data, transforming it into new data. The reduce part of the model combines those outputs into something understandable.

You could use this system to calculate the total amount of money in a bank by summing all clients accounts.

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

Similar to MapReduce, Spark RDDs are used to analyze huge amounts of data. However, this model instead of using files, it maintains the data in memory.

This makes Spark RDD much faster and writing jobs is cleaner and quicker. But, this makes Spark very resource intensive.

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

Supervised Learning is the type of machine learning that predicts with examples. You feed the algorithm some data and then label that data with the correct answer. Then the predictions it makes are based the labels that were provided.

For example, you can teach the algorithm to predict 'BUY' or 'SELL' with current buying or selling data, then ask it to predict whether to buy or sell on another stock.

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

In Unsupervised Learning, the algorithm decides how to structure the data without indications on whether it's doing it correctly or not.

The most common way of doing this is clustering; sorting the data into clusters depending on some features and how they relate to one another. When a prediction is made, the incoming data is analyzed to see how well it fits into a cluster.

You could use this system to detect anomalies in systems or if an attacker is impersonating someone.