University of Puerto Rico
Mayagüez Campus

Department of Computer Science and Engineering

CIIC 5995/8995: Big Data Analytics
Final Exam

Name: Carlos Theran TOTAL: /100

Student ID: 502-114736

Section: 8995 **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 es una tecnica de procesamiento de grandes coleccion de datos bajo un modelo de programacion distribuida sobre un grupo de computadoras "commodity". Su nombre se debe a sus dos principales funciones Map y Reduce, en las cuales Map se encarga de mapear cada elemento del conjunto de datos con un valor retornando una lista de pares. Reduce agrupa todo los pares con la misma clave, genernado grupos por cada una de las diferentes claves.

Ejemplo.
Si se tiene una base de datos enorme, en la cual se ha guardo una encuesta la cual tiene datos de una persona como nombres y apellido, ocupacion, direccion de residencia (barrio, ciudad, calle, pais), fecha de nacimiento entre otros. Si se desea conocer el numero de personas con la misma ocupacion o personas que sean solo estudiantes en un barrio especifico se hace uso de MapReduce.

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

Spark RDDs: Es una coleccion de objetos immutables (read-only) cada conjuto de datos en un RDD es divido entre particiones logicas sobre diferentes nodos del cluster. Asi los RDD son visto como coleccion de objetos que pueden ser operados en paralelo.

Es de saber que MapReduce es bastante usado para el proceso de grande cantidad de datos, lamentablemente la unica forma de reusar la data entre computos es guardandola en un sistema de almacenamiento externo (HDFS) lo cual convierte el proceso bastante lento para aquellas aplicaciones iterativas. La mayoria de aplicaciones en hadoop gastan aproximada mente el 90% del tiempo haciendo un HDFS read-write operation. Ahora Spark RDD soporta "in-memory proccesing computation", almacena la data en memoria para su procesamiento esto permite que sea de 10 a 100 veces mas rapido.

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.

Es una tecnica de machine learning en la cual se provee una data de entrenamiento, la cual contiene un "input" y su respectivo "output". Con la data de entrenamiento se entrena el algoritmo para que mape su "input data" con su respectivo "outpu data" y de esta manera clasificandola entre clases. De esta forma el algoritmo esta preparado para clasificar nueva data entre las clases ya antes creadas.

Ejemplo.
Un exelente ejemplo son los dispositivos electronicos que requieren reconcer la huella dactilar de su propietario para que este se pueda desbloquear. Cuando se obtiene este dispositivo por primera vez usted debe proveer su huella para que el celular genere la data de entrenamiento, para posteriormente saber clasificar los pixeles de la imagen de su huella dactilar y asi aceptar o rechazar aquellas imagenes de huellas que no sea bien clasificada dada su dato de entrenamiento.

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.

Esta tecnica a diferencia de la supervisada solo se provee "input data" es decir, no se conoce de antemano el "output" de esta data. Esta tecnica busca una estructura o relacion entre la data suministrada (input).

Ejemplo.

Un ejemplo de este tipo de tecnica seria implementar k-nn o clustering para segmentacion de una imagen.