University of Puerto Rico
Mayagüez Campus

Department of Computer Science and Engineering

CIIC 5995/8995: Big Data Analytics
Final Exam

Name: Cristian Camilo Garzón Alfonso TOTAL: /100

Student ID: 502-16-1887

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 es un método de programación que nos ayuda a procesar y generar grandes volúmenes de datos. El usuario especifica una función de mapeo la cual se encarga de procesar un dupla compuesta <llave,valor> muchas veces la llave no se usa en esta fase. El resultado de esta primera fase es un conjunto de duplas <llave, valor> intermedias, en la fase de reducir une todos los valores que tienen la misma llave y aplica una función sobre estos valores, bien sea una sumatoria u otra función de agregación. Este tipo de programación corre en modo paralelo por el clúster que se esté ejecutando.

Ejemplos:
Conteo de palabras, se quiere saber cuántas veces se repite una palabra dentro de un texto.
Obtener las N palabras que más se repiten en un texto.
Sumatoria de números, se quiere saber el total de ventas de los productos de una tienda.

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

El API de programación que provee SPARK, se manipula por medio de los Resilient Distributed Datasets (RDDs). Cada RDD es una colección de objetos del lenguaje en el que se esté programando Scala, Java o Python, que están particionados y repartidos en los diferentes nodos del clúster. Estos RDDs tienen funciones de Map, Reduce, Filter, entre otras. Los RDDs son “Lazy”, esto quiere decir que no ejecuta las funciones que se le piden hasta que haya una operación que lo obligue a hacer el computo como un count, collect, entre otras. Spark corre sus procesos en memoria RAM lo que le permite tener una ventaja de rendimiento a la hora de compararse con una tarea de MapReduce normal que corre en disco duro.

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.

En el aprendizaje supervisado se da un conjunto de datos que ya tiene identificado el resultado correcto que debe arrojar. Se pueden categorizar en problemas de regresión y clasificación. En un problema de regresión los resultados se predicen usando las variables de entrada como parámetros para una función continua que la define. En un problema de clasificación lo que se trata de predecir es a que categoría específicamente pertenece, es decir entrega un valor discreto.

Ejemplos:
Predecir el valor de un auto, si se tiene el historial de ventas de autos de los últimos n años.
Predecir si un tumor es maligno o benigno.
Predecir el valor de cualquier producto en un almacén de cadena.
Predecir la edad de una persona, por medio de una foto.

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.

En el aprendizaje no supervisado podemos abordar el problema sin necesidad de conocer como deberían ser nuestros resultados, este tipo de aprendizaje nos da grupos de datos que están agrupados por ciertas características que son comunes para los elementos del grupo. Aquí no se tiene una retroalimentación, es decir, no podemos juzgar y decir si la respuesta es correcta o incorrecta.

Ejemplos:
Agrupar los clientes de una tienda por las cosas que compra y consume.
Tomar un gran número de registros médicos y clasificar los pacientes en diabéticos, hipertensos, con problemas del corazón, etc.
Identificar voces de grabaciones de sonidos.