University of Puerto Rico Mayagüez Campus

Department of Computer Science and Engineering

CIIC 5995/8995: Big Data Analytics
Final Exam

Name: Danny G Villanueva Vega TOTAL: /100

Student ID: 502-16-4863

Section: 116 **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.

Es un modelo simple para cluster computing, se divide en dos procesos: Map, para el procesamiento inicial de los datos (filtering, transformation) y reduce para el procesamiento final (accumulation, aggregation). Se puede usar equipos cómodos y accesibles, es tolerante a fallas, altamente escalable y tiene APIs disponibles en Java, Python, C++.

Los casos de uso de MapReduce son los siguientes: Data analytics, Crawling, Full-text indexing, Reputacion systems, Data mining.

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.

Un RDD es una abstracción de una colección de datos distribuidos en varios nodos, representa una colección de elementos inmutables y divididos que pueden ser operados en paralelo.

Es una alternativa a MapReduce, el procesamiento de los datos en spark se realiza en memoria, incrementando el rendimiento en gran medida, sigue el modelo Directed Acyclic Graph (DAG), y los datos los mantiene en memoria, de este modo no materializa los resultados intermedios en el disco.

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.

El aprendizaje supervisado son algoritmos que aprenden a realizar predicciones a partir de un conjunto de ejemplos de entrenamiento, se usan los algoritmos de regresión, clasificación, y redes neuronales.

Se puede usar en la bioinformática, la informática química, reconocimiento de escritura a mano, reconocimiento de objetos en visión computacional, detección de spam, patrones de reconocimiento, reconocimiento de voz, etc.

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.

Aprendizaje no supervisado, es una categoría especializada de machine learning que usa algoritmos que aprenden la estructura de los datos por si mismos, es decir infieren por si mismos la estructura. Usa algoritmos como clustering, outlier detection, etc.

La detección basada en comportamientos en la seguridad de la red es una buena área de aplicación, esto es porque la cantidad de datos para un analista humano de seguridad es imposible (medido en terabytes por día), para revisar y encontrar patrones y anomalías, el aprendizaje de las maquinas para la industria de la seguridad es su capacidad para detectar ataques avanzados y desconocidos, el papel del aprendizaje automático es crear perfiles continuos para usuarios y dispositivos y luego encontrar anomalías significativas.