University of Puerto Rico
Mayagüez Campus

Department of Computer Science and Engineering

CIIC 5995/8995: Big Data Analytics
Final Exam

Name: Mario Orbegoso Villanueva

TOTAL: /100

Student ID: 801-12-5755

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.

MapReduce is a programming paradigm that aims to tackle computing tasks that involve processing a lot of data in a scalable manner that runs on a cluster of computers. This programming paradigm divides a task in two phases, map and reduce, and executes each of them in a parallel fashion. In this parallel execution instead moving the data to execute computations, it does the opposite, it brings the computation to the data. This is possible since the data is actually not in one central location, but it is distributed throughout a cluster in separate machines with compute power.

In the map phase:
It reads data structured in a key-value structure (k1, v1) and produces a mapping to a different key-value pair (k2, v2) that will be passed down the reduce phase. This phase is used for cleaning up the data and structure it in a more standard and readable way according to the needs of the problem it wants to solve.

In the reduce phase:
It receives the aggregated output of the map phase (k2, v2) and applies the final computation on these values.

Solves the following problems that involve big data:
- Statistical methods
- Machine learning
- Social media analytics

… in essence, anything which involves big amounts of data. i.e. in the realms of Petabytes

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 stands for: Resilient Distributed Datasets. It is the base data structure of which Spark is built upon, it is a representation of data that is actually distributed across a cluster of computers and stored in memory. Formally, an RDD is read-only data that can be created by using deterministic operations on other RDD or stored data like HDFS.

The major difference between Hadoop MapReduce and Spark is that of the method of how they store the intermediate results of the data. While Hadoop MapReduce's stores it in disk and produces more overhead for reading and writing to disk, Spark's approach saves and computes everything in memory which produces faster results. This allows Spark to perform in real-time and opens up a whole category of computing problems that can be solved in this manner.

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 one of the two broad specializations of Machine Learning.
It is when one hands the algorithm a set of example datasets for it to train. The example dataset consists of a series of records which contain an input (typically a vector) and a desired/correct output value. The goal is for the algorithm to produce an inferred mapping function that can later be used to classify other similar (but not the same) datasets to the one that was used to train.

Examples:

- Regression
    - Twitter Sentiment Analysis
- Classification
    - Facial Recognition
    - Object Recognition

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 one of the two broad specializations of Machine Learning. Is when an algorithm gets handed a dataset and expects the algorithm to learn on it's own. The data set consist of records of only input and no output/desired value, there is no correct output/answer. The goal is to let the algorithm itself discover or find relations and find a structure within that dataset by its own.

Examples:
- Clustering
- Outlier Detection