

University of Puerto Rico  
Mayagüez Campus

Department of Computer Science and Engineering

CIIC 5995/8995: Big Data Analytics  
Exam 1 - Part II: Programming Exercises

Name: \_\_\_\_\_

TOTAL:     /60

Student ID: \_\_\_\_\_

Section: \_\_\_\_\_

March 30, 2017

In this part of the exam you will write code in Java, Hive-QL or Python.

- I) **(20 pts) Problem I: MapReduce - *Folder mapreduce***: Consider the collection of tweets from project 1. Write MapReduce code to find the set of all tweet ids on which the following keywords appear.

- (a) MAGA
- (b) Dictator
- (c) Impeach
- (d) Drain
- (e) Swamp
- (f) Change

For example, the output line for MAGA should look something like this:

MAGA, 12353534, 124923423, 9584934, 859584400

In this case, we have the keyword, followed by the list of tweet ids for the tweets in which MAGA appears.

Tasks:

- (1) **(10 pts)** Write the Map function as:

```
public class KeywordToTweetsMapper extends
    Mapper<LongWritable, Text, Text, Text> {
    @Override
    public void map(LongWritable key, Text value, Context context)
        throws IOException, InterruptedException {
```

(2) (10 pts) Write the reduce function as :

```
public class KeywordToTweetsReducer extends
    Reducer<Text, Text, Text, Text> {
    @Override
    protected void reduce(Text key, Iterable<Text> values, Context context)
        throws IOException, InterruptedException {
```

II) (20 pts) **Problem II: Hive - *Folder hive***: Consider the data provided in files `escuelasPR.csv`, and `studentsPR.csv`. The schema for each file is found in files `escuelasPR_schema.txt` and `studentsPR_schema.txt`.

Provide Hive-QL queries for the following tasks:

- 1) (5 pts) Find the total number of schools in each region, and city.
- 2) (5 pts) Find the total number of students per school.
- 3) (5 pts) Find all the students that go to school in the city of Ponce or in Cabo Rojo.
- 4) (5 pts) Find the total number of students per region and city.

Store you solution on a file name `hive.sql`.

III) (20 pts) **Problem III: Spark - *Folder spark***: Consider the data provided in files `escuelasPR.csv`, and `studentsPR.csv`. The schema for each file is found in files `escuelasPR_schema.txt` and `studentsPR_schema.txt`.

Provide the following python code to complete the following tasks in pyspark:

- 1) (10 pts) Find all the female students enrolled in school with id 71381. **NOTE: Provide your solution on a file named p3a.py**
- 2) (10 pts) Find all the schools in the city of Ponce or in Cabo Rojo or in Dorado. **NOTE: Provide your solution on a file named p3b.py**