COMP 3610 - Big Data Analytics

The University of the West Indies, St. Augustine

Assignment 2

Date Distributed: March 10th, 2020 **Date Due:** March 21st, 2020 (11:59pm)

DataSet

The dataset for this assignment can be found here.

Additional resources for understanding the main concepts with use cases:

- Data Algorithms Recipes for Scaling Up with Hadoop and Spark
- Spark Practice
- Complete Guide on DataFrame Operations in PySpark
- Spark Transformations in Python

Objectives

The objectives of this assignment:

- Create RDD (Resilient Distributed Dataset) from a list of numbers, tuples and data loaded from files
- Manipulating RDDs to carry out specific tasks
- Converting RDDs to a PySpark's DataFrame
- Carrying out analysis on RDDs
- Converting PySpark DataFrames to RDDs
- Carrying out analysis on PySpark DataFrames

Part A

1. Convert the following sentence to a python tuple list of letters and their frequency as their appear. Ignore all non-alpha numeric characters. **5 marks**

   Sentence:

   "The quick brown fox jumps over the lazy dog and the fox was very happy"

   Tuple List:

   [('h', 1), ('e', 1), ('t', 1), ('q', 1), ('i', 1), ('c', 1), ('u', 1), ('k', 1), ('r', 1), ('b', 1), ('o', 1), ('w', 1), ('n', 1), ('x', 1), ('o', 1), ('f', 1), ('u', 1), ('s', 1), ('j', 1), ('m', 1), ('p', 1), ('r', 1), ('e', 1), ('o', 1), ('v', 1), ('h', 1), ('e', 1), ('t', 1), ('a', 1), ('y', 1), ('z', 1), ('l', 1), ('o', 1), ('d', 1), ('g', 1), ('a', 1), ('d', 1), ('n', 1), ('h', 1), ('e', 1), ('t', 1), ('x', 1), ('o', 1), ('f', 1), ('a', 1), ('s', 1), ('w', 1), ('y', 1), ('r', 1), ('e', 1), ('v', 1), ('a', 1), ('h', 1), ('y', 1), ('p', 2)]

2. Create a PySpark Context **1 marks**

3. Convert the list of tuples to a PySpark RDD. **2 marks**

4. Using the methods of PySpark RDD display the letter count. **3 marks**

   Sample:

   [('a', 4), ('e', 5), ('i', 1), ('m', 1), ('q', 1)]

5. Using the methods of PySpark RDD display the letter and number of times they appear in each word in the sentence. **3 marks**

   Sample:

   [('a', [1, 1, 1, 1]), ('e', [1, 1, 1, 1, 1]), ('i', [1]), ('m', [1])]

Part B

If you are a frequent user of Amazon.com, you are probably familiar with the lists of related products (books, CDs, etc.) the site features to help customers find what they are looking for. Amazon.com presents several such lists on every page, including “Frequently Bought Together” and “Customers Who Bought This Item Also Bought.” These features have roots and solutions in recommendation engines and systems.

In this task students are asked to determine:

- Customers who bought this item also bought
- Most popular items

1. Create a sql context from PySpark SQLContext. **1 marks**

2. Load the Amazon Review Dataset into a PySpark RDD, ensure that each row is properly separated and the headers are matched to their respective columns. **5 marks**

3. Convert the rdd to a PySpark DataFrame. **1 marks**

4. Using the dataframe from the above question show the top 20 bought products. **5 marks**

5. Using the datafrme from the previous question show the top 20 users and the products that they purchased. **4 marks**

6. Create a RDD of tuples from the dataframe with only 2 columns ‘product’ and ‘username’ in that order. **1 marks**

7. Using methods from PySpark’s RDD object e.g. groupByKey, map, reduceByKey derive the top 20 products. **5 marks**

   Sample:

   [(u'Amazon Tap - Alexa-Enabled Portable Bluetooth Speaker', 542),

   (u'Amazon Fire TV', 166),

   (u'Amazon Premium Headphones', 133),

   (u'Fire HD 6 Tablet', 87),

   (u'Kindle Fire HDX 7', 53)]

8. Create another RDD of tuples from the dataframe with the columns ‘username’ and ‘product’ in that order. **1 marks**

9. Using methods form PySparks’s RDD object product the top 10 customers who purchased the most items. The top 10 list must show the username and a list of all the items they bought with the number of that item they bought. **8 marks**

   Sample:

   [(u'A. Younan', ({u'Amazon Premium Headphones': 59}, 59)),

   (u'William Hardin',

   ({u'Amazon Fire TV': 16,

   u'Certified Refurbished Amazon Fire TV (Previous Generation - 1st)': 12,

   u'Fire HD 6 Tablet': 30},

   58)),

   (u'Andrew', ({u'Amazon Premium Headphones': 43}, 43))]

Assignment Requirements:

For this assignment students are required to install Spark and PySpark.

- Installing Spark and PySpark for windows
- Installing Spark and PySpark for mac

Submission Details:

1. Ensure that your name and ID and present in the notebook.
2. Export/download your file from Jupyter notebook.

3. Email your file to ssooklal27@gmail.com. Put as the title of the email COMP 3610 A2.
4. There would be a 10% penalty per day for late submissions, up to 5 days.