COMP 6940
Assignment 2

Date Due: Sunday 18th March, 11:59pm

Introduction

The aim of this project is to provide students with some practical experience in analyzing data to come up with interesting results that enable more informed decision making. The project seeks to utilize real data to provide students with the experience of extracting information that actually makes sense and is directly applicable to their experiences.

The focus is for students to grasp the necessary skills to be able to manipulate data using Apache Spark Cluster Computing.

DataSet:
The dataset for this assignment can be found here.

Additional Resources for Understanding the concept of MapReduce 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

Duration: 2 weeks

Assignment Total: 45 marks

Task A:

1. Convert the following sentence into a python tuple list of letters and the frequency of which each letter appears in the current word. Ignore all non-alpha numeric characters.

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

   Sample 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)]

   5 marks

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

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

4. Using the methods of PySpark RDD display the letter count.
   Sample:
   [('a', 4), ('e', 5), ('i', 1), ('m', 1), ('q', 1)] **3 marks**

5. Using the methods of PySpark RDD display the letter and number of times they appear in each word in the sentence.
   Sample:
   [('a', [1, 1, 1, 1]), ('e', [1, 1, 1, 1, 1]), ('i', [1]), ('m', [1])] **3 marks**

Task B:

If you are a frequent user of Amazon.com, you are probably familiar with the lists of related products (books, CDs, etc.) which 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 will be 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 into a PySpark DataFrame. **1 marks**

4. Using the dataframe from question 3 show the top 20 bought products. **5 marks**

5. Using the dataframe from question 3 show the top 20 users and the products that they purchased. **4 marks**

6. Create a RDD of tuples from the dataframe from question 3 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.
   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)] **5 marks**

8. Create another RDD of tuples from the dataframe from question 3 with the columns 'username' and 'product' in that order. **1 marks**

9. Using methods from PySparks's RDD object, produce the top 10 customers who purchased the most items. The top 10 list must show the usernames and a list of all the items each person bought. Each item should have an associated quantity value representing the amount of the item which was purchased by the customer.
   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))] **8 marks**

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:
Students should submit either the following:

- An exported Jupyter notebook of the assignment

or

- A python script file with the assignments and appropriate comments

Email submissions to: nchamansingh@gmail.com with the following subject format:
<StudentID> COMP6940 Assignment 2