INFS 4020 - Big Data Concepts

Practical Test 3 (SP5 2022)

Due: By 11PM on Friday 28 October

General Instructions

- This test is worth 10% of your final grade and it is due no later than 11pm on Friday 28 October.
- The test will be marked out of 20.
- You will need to submit your work via learnonline in **zip** format.
- You will need to **STRICTLY** follow the Submission Instructions.

Assessment Tasks

In this assessment you are required to answer five questions using Spark. For each question you will need to write code which uses the appropriate transformations and actions. You are free to use RDD and/or DataFrame to answer the questions.

Our main input file for this assessment is called **DataCoSupplyChainDataset.csv**, which contains supply chain data used by a company called DataCo Global. There is a second file provided, called **DescriptionDataCoSupplyChain.csv**, which describes the columns in the main dataset.

To prepare the data, first, create a directory for this assessment called **test3** within the **/home/prac/** directory as we normally have in practicals.

Then, copy the data file into your **/home/prac/test3/input** folder.

Lastly, create the **/home/prac/test3/src** directory for your Python code and **/home/prac/test3/output** directory for writing the result file. The result file is expected to be generated in the directory specified in the template file (/home/prac/test3/output/result.txt). From here you should be able to follow the directions in Practical 4&5 to write and run your PySpark programs.

The program will be run using the following command when marking:

```
$ spark-submit /home/prac/test3/src/test3_solutions.py
```

The command should write the output of your program in the **/home/prac/test3/output** directory as specified in the template file. Your program will be marked by comparing the result from your program to the correct answer. Rounding the result is not required, but you will not lose marks if you do so.

You are provided with a template for the **test3_solutions_template.py** file. The easiest way to complete the assessment and fill in the template will be to work in the terminal, then copy and paste your code to the .py file.

Q1. Load the data, convert to DataFrame and apply appropriate column names and variable types.

Q2. Determine what proportion of all transactions is attributed to each customer segment in the dataset i.e. Consumer = x%, Corporate = y% etc.
This question uses the **Customer Segment** field.

Q3. Determine which three products had the least amount of sales in revenue.
This question uses the **Order Item Total** and **Product Name** fields.

Q4. For each transaction type, determine the average item cost before discount.
(Tip. use *Total Cost* / *Total Quantity* to calculate average)
This question uses the **Type**, **Order Item Product Price** and **Order Item Quantity** fields.

Q5. What is the first name of the most regular customer in EE. UU.? (Repeat transactions by the same customer, with the same Id, should not count as separate customers).
This question uses the **Customer Country**, **Customer Fname** and **Customer Id** fields.

Submission Instructions

You should submit your Spark program file in a **zip** file with the name below:

- **test3_solutions.py**, this should be the template provided filled with the appropriate code.

Make sure to comment your code sufficiently, this will be included in your final mark. Good programmers are good commenters too, your code should be able to be read by a stranger who wants to use it, or by yourself in a year's time.

Once finished, zip the program file and upload your zip file to learnonline.

Distribution of marks

Q1 – 3 marks

Q2 – 4 marks

Q3 – 3 marks

Q4 – 5 marks

Q5 – 4 marks

Overall code presentation – 1 mark (good variable naming and comments)

Total of 20 marks