**Database Project (SWE3033) (Fall 2023)**

**Homework #4 (50pts, Due date: 10/11)**

**Student ID**: 2020315798

**Student Name**: Choi Jin Woo

**Instruction:** In this homework, we provide you with a jupyter notebook file (DBP\_Homework4.ipynb). You should follow the instructions in these documents carefully.

**Submit two files as follows:**

- DBP\_Homework4\_StudentID.zip

- DBP\_Homework4\_StudentID.ipynb

- DBP\_Homework4\_StudentID.pdf

**1. [10pts]** Calculate the visit frequency for each user to the places James and Mary visited.

1. Places that James visited:

- [‘E-mart’, ‘Starbucks’, ‘GS25’, ‘Starbucks’, ‘HomePlus’, ‘CU’]

1. Places that Mary visited:

- [‘Starbucks’, ‘E-mart’, ‘Starbucks’, ‘LotteMart’, ‘LotteMart’]

**[Answer]**

Enter your code and result here. You must show your result (captured image).

|  |
| --- |
|  |

**2. [20pts]** Count the number of words in the given data using the following two operations and explain the difference between the two operations.

***Data:***

[(‘odd’, 1), (‘odd’, 1), (‘even’, 1), (‘odd’, 1), (‘even’, 1), (‘odd’, 1), (‘odd’, 1), (‘even’, 1), (‘even’, 1)]

1. groupByKey()
2. reduceByKey()
3. Explain the difference between the two operations.

**[Answer]**

|  |  |
| --- | --- |
| **a)** | [('even', 4), ('odd', 5)] |
| **b)** | [('even', 4), ('odd', 5)] |
| **c)** | [groupByKey()]  This operation groups the data by key and produces an iterable of values  For the intermediate step, it reaches [('odd', [1, 1, 1, 1, 1, 1]), ('even', [1, 1, 1, 1])] and then gets the final result to [('odd', 5), ('even', 4)]  [reduceByKey()]  This operation combines the values for each key using a reduction function and provides the word count as the result.  For the intermediate step, it reaches directly to [('odd', 5), ('even', 4)] |

Enter your code and result here. You must show your result (captured image).

|  |
| --- |
|  |

**3.** **[20pts]** The following data represents the songs Mary and James have listened to and the play counts. Answer the following three questions.

***Data:*** key-value data in (music, # of plays) format

- James: [(‘Thriller’, 30), (‘Everybody’, 34), (‘Everybody’, 30), (‘Billie\_Jean’, 2)]

- Mary: [(‘Thriller’, 20), (‘Sorry’, 23), (‘Sorry’, 3), (‘Billie\_Jean’, 5)]

1. For each user, calculate the number of times each song has been listened to, store it in a new RDD. (HINT: reduceByKey())
2. Create a new RDD containing songs that both users have listened to and their respective play counts. (HINT: join())
3. Calculate the total number of music plays that James and Mary have played in common.

**[Answer]**

Enter your code and result here. You must show your result (captured image).

|  |
| --- |
| a.    b.    c. |