IDS 561 Homework 3

Due date: 03/30/2021 Tuesday 3:30pm (before class)

In this homework, you will perform Spark DataFrame/SQL Transformations and Actions.

Dataset

You will use the same dataset of HW2. Please download the dataset from this link:

https://www.dropbox.com/s/64lm3yxcfkb0hl8/Amazon_Responded_Oct05.csv?dl=0

The ***Amazon_Responded_Oct05.csv*** contains information of 400K tweets. There are 3 columns that you will use for this assignment.

Columns	Meaning
tweet_created_at	When was the tweet created
user_screen_name	User screen name
user_id_str	User id

Task

Step 1: Find out the users who are active in at least five listed days (i.e., created posts in at least 5 days) in ***Amazon_Responded_Oct05.csv*** and save their "user_screen_name" and "user_id_str" in the dataframe "***daily_active_users***" (see below). Report how many active users you find.

daily_active_users

user_screen_name	user_id_str
AmazonHelp	85741735
…	…

Step 2: A company would like to conduct an A/B test on Twitter. The ***experiment.txt*** file includes the user_id_str they selected as potential experiment targets. Please create a dataframe "***experiment_user***" to document the selected user id and whether they are active users (join the dataframe from step 1). For example:

experiment_user

user_id_str	whether_active
85741735	yes
…	…

Then calculate the percentage of active user and print out the result.

Step 3: In homework 2, you have already known how to join 2 tables in spark. Now you are going to perform a 3-table join task.

The company provided their revised experiment target list in ***final_experiment.csv*** file. Compared with the former experiment.txt file, they removed several users and added a new column "info" to indicate whether the user is female (F) or male (M). However, they are still missing some information.

① Please help them fill in the remaining columns by joining the dataframes you got from step 1&2 together and save the result in a dataframe "***final_experiment***". (Note: For inactive users that cannot be found in "daily_active_users", you can leave their "user_screen_name" blank or fill with "Not found")

final_experiment

user_id_str	info	whether_active	user_screen_name
85741735	F	yes	AmazonHelp
12345678	M	No	
…		…	

② Please describe your join steps briefly. For example, which joins (inner, outer, etc.) did you use.

What to submit (one submission per group)

You need to submit: a Python file and three .csv files.

Python file: Your code. Please add comments to make it readable.

CSV files: Export the three dataframes and save them as .csv files (daily_active_users.csv, experiment_user.csv, final_experiment.csv).