## Dataset

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

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

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

## Task

*Step 1*: Create a dataframe "daily_active_users". 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. For example:

***daily_active_users***

| user_screen_name | |
|---|---|
| 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***

| | Whether_active |
|---|---|
| 85741735 | yes |

| | |
|---|---|
| … | … |

Then, I calculated the percentage of active users and printed out the result.

*Step 3*: Next, I performed a 3-table join task.

To help the company prepare the data, I have selected the records (all columns) in ***Amazon_Responded_Oct05.csv*** when a user_id_str is included in all the 2 dataframes. For example, if the user_id_str from ***Amazon_Responded_Oct05.csv*** cannot be found in daily_active_user and experiment_user, I skipped.

## Output

Saved the result in a dataframe and then export it as ***Amazon_new.csv***