In this , you will perform Spark RDD Operations - Transformations and Actions. You are required to write the raw code for RDD (not using any dataframe API).

**Dataset**

Please download the dataset from this link:

[https://www.dropbox.com/s/64lm3yxcfkb0hl8/Amazon\_Responded\_Oct05.csv?dl=0](%22)

The ***Amazon\_Responded\_Oct05.csv*** contains information of 400K tweets. There are 6 columns that you will use for this assignment.

|  |  |
| --- | --- |
| **Columns** | **Meaning** |
| id\_str | tweet ID |
| tweet\_created\_at | when was the tweet created |
| user\_verified | whether the user is verified (TRUE or FALSE) |
| favorite\_count | how many times the tweet is favorited |
| retweet\_count | how many times the tweet is retweeted |
| text\_ | text content of the tweet |

**Task 1**

*Step 1*: Remove the records where “user\_verified” is “FALSE”.

*Step 2*: For the remaining records (“user\_verified” is “TRUE”), group by created date, and count the number of tweets for each date.

Example: If “tweet\_created\_at” is “Tue Nov 01 01:57:25 +0000 2016”, the created date is “Nov 01”.

*Step 3*: For the date with highest number of tweets (you can figure it out from step 2), calculate the sum of “favorite\_count” and “retweet\_count” for each tweet on that day. Then report the text content (“text\_”) of the top 100 tweets with highest sum. Count the word frequency of the 100 tweets and report the result (Note that data cleaning steps before wordcount can be done outside of Spark operations).

**Task 2**

You will use ***find\_text.csv*** for this task. There are two columns in this document: “id\_str” and “text”. The second column is empty. Please find out the text content of each tweet according to “id\_str” joining ***Amazon\_Responded\_Oct05.csv*** and fill in the “text” column.

Note: If a tweet ID appears in multiple records, just report the text content from one of them.

**What to submit**

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

**Python file (.ipynb)**: Your code. Please mark the task number and step number clearly in your code and **add comments** to make it readable.

**CSV file**: Fill in the text column in find\_text.csv and submit it.

**PDF file**: Your brief answer to the following questions:

Q1: how many records that were created by verified users (“user\_verified” is “TRUE”)?

Q2: which day has the highest number of tweets?

Q3: word count output in task 1 step 3. The format should look like this:

|  |  |
| --- | --- |
| **Word** | **Frequency** |
| apple | 123 |
| banana | 45 |
| … | … |