This project focuses on performing Spark RDD Operations - Transformations and Actions.

Dataset

The ***Amazon_Responded_Oct05.csv*** contains information of 400K tweets. There are 6 columns that you will use for this assignment.

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

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 the 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 the highest sum. Count the word frequency of the 100 tweets and report the result.

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.

Output

The .csv file contains words and their frequency from the input file