# Task A: Shell Commands

**Question Description**:

Are you a fan of sports? Have you paid attention to the Olympic games held in Tokyo - Japan

earlier this year? In this task, you are required to explore and wrangle the data in the file

“Olympics\_tweets.csv”, which contains tweets related to the Tokyo Olympic 2021 that were

collected from Twitter. The file contains different variables to describe the collected tweets,

e.g., the id of a tweet, the text content of a tweet, the screen name of a user who posted a

tweet (i.e., user\_screen\_name), and so on. Please note that you are only allowed to use shell

commands as you would run in Linux shell, Mac terminal, or Cygwin, to tackle this task.

Using other utilities or tools such as PowerShell is NOT allowed.

1. The first step is to explore the contents of the file. Please write commands to display the first 15 rows and the last 15 rows of the data file.

**First 15 rows:**

**First way:**

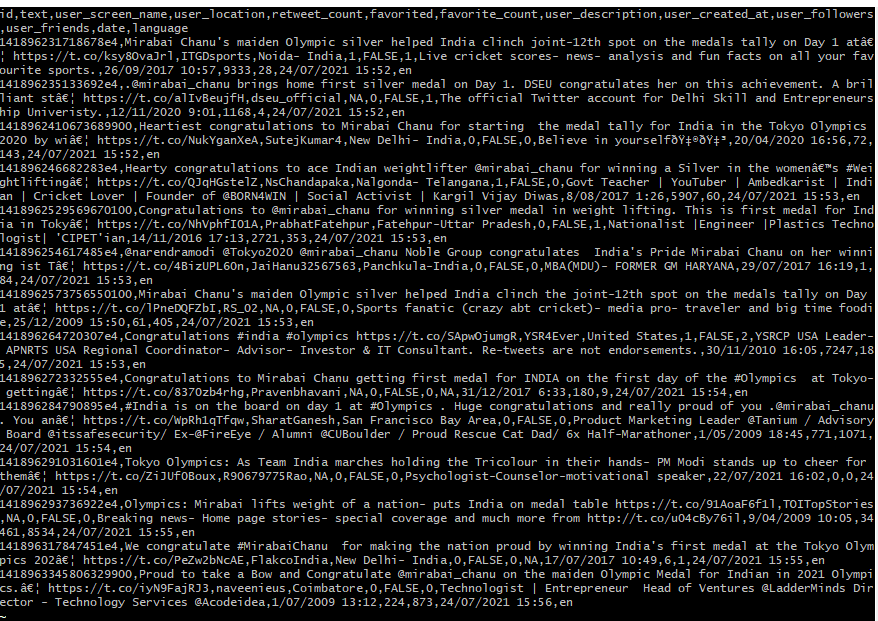
Code Explanation:

Open the csv file | split by ‘,’ and select all column between 1 to 15 | open in editor

Code:

cat Olympics\_tweets.csv | awk -F',' 'NR>=1 && NR<= 15' | less

Output:



**Second way:**

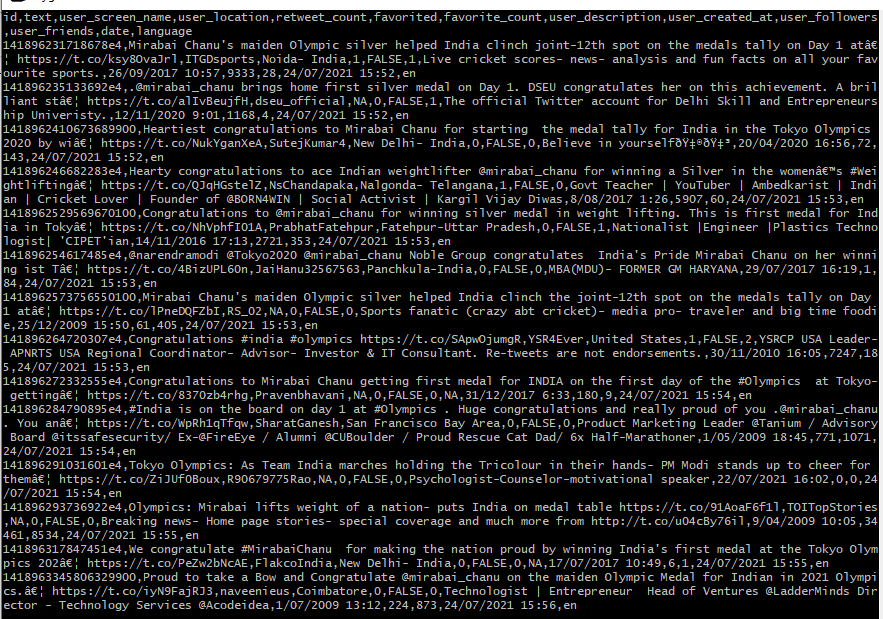
Code Explanation:

Open the csv file | select first 15 rows using head command | open in a editor

Code:

cat Olympics\_tweets.csv | head -n15 | less

Output:



**Last 15 rows:**

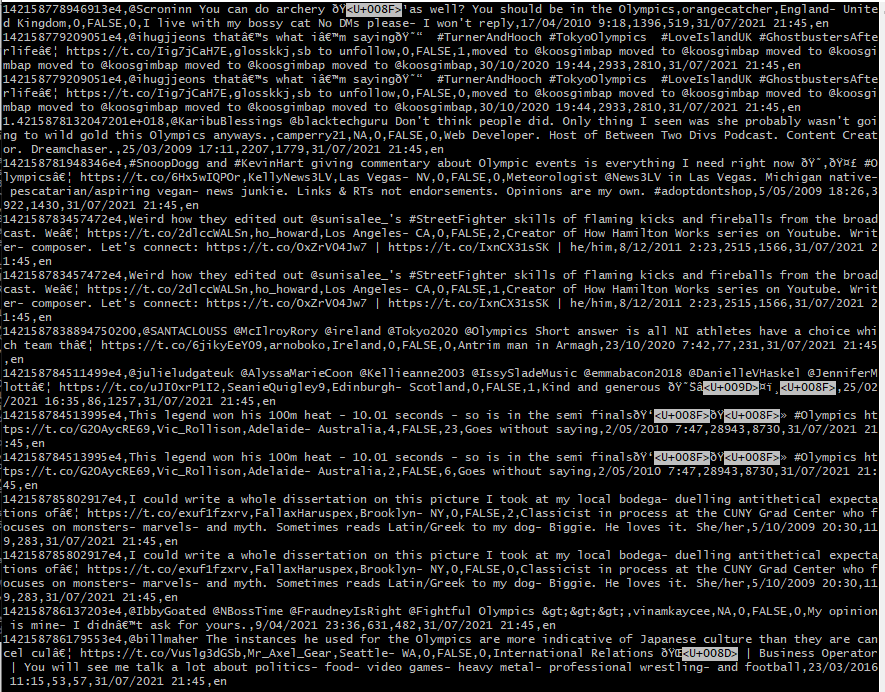
Code Explanation:

Open the csv file | select bottom 15 rows | open in a editor

Code:

cat Olympics\_tweets.csv | tail -n15 | less

Output:



1. Write commands to display the size of the file and number of lines inside the file

Code Explanation:

Display the size of the file and number of line inside the file

Code:

wc Olympics\_tweets.csv

output:

lines, word, byte counts

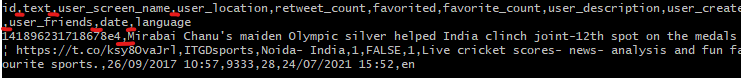


1. What is the field separator in the file? Write a command to display the number of variables (i.e., the columns) in the data and a command to display the column names.

**What is the field separator?**

The field operator is ‘,’

Example:



**Number of variables**

Code Explanation:

display the number of variables

Code:

awk -F, '{ print NF; exit }' Olympics\_tweets.csv

Output:



**Columns names**

Code Explanation:

The first line is the column names and head -n 1 means display the first line

Code:

head -n 1 Olympics\_tweets.csv

Output:



1. Write commands to count and then remove lines with an id that is not a number of 19 digits long, i.e., id values that contain anything other than numbers OR are of a length more/less than 19. Store the filtered set in a file named “filtered\_tweets\_1.csv” and use this file for the following questions in Task A.

Code Explanation:

Split by ‘,’ if values’ length in the first line is equal to 19 or it’s the column names line, print. Save all selected values in a new file named filtered\_tweets\_1.csv | open the new file | count lines and words and bytes count.

Code:

awk -F',' '{if (length($1) == 19 || NR==1) print }' Olympics\_tweets.csv > filtered\_tweets\_1.csv |cat filtered\_tweets\_1.csv| head -n -1 | wc

Output:

Lines, words, bytes count



1. What is the date range of the tweets? Please note that the file is not guaranteed to be sorted and Nulls (NA and empty values) should not be considered.

Code Explanation:

Open the csv file | split by ‘,’ if column 12 is not null value and its not the first line, print column 12 | sort the column | print the first and last dates.

Code:

cat filtered\_tweets\_1.csv | awk -F',' '{if($12!=NULL && NR > 1) print $12}' | sort | (head -n1 && tail -n1)

output:



1. When was the first mention of the term “Japan” in the dataset? What about “Australia”? Please note that the first mention of a term refers to the creation time of the tweet that contained the specific term for the first time in the dataset and the terms to be searched are case sensitive.

**Japan**:

Code Explanation:

Open the csv file | find the word Japan | find the first column | print column 12

Code:

$ cat filtered\_tweets\_1.csv | grep 'Japan' | head -n1 | awk -F"," '{print $12}'

output:



Australia:

Code Explanation:

Open the csv file | find the word Australia | select first line | print column 12

Code:

cat filtered\_tweets\_1.csv | grep 'Australia' | head -n1 | awk -F"," '{print $12}'

Output:



1. How many unique “user\_screen\_name” values are there in the dataset? Can you list the top 10 most frequent user\_screen\_name in the dataset?

Total unique values:

Code Explanation:

Open the csv file | split by ‘,’ print third column | remove first line | sort values | find unique values | sort values | count total rows

Code:

cat filtered\_tweets\_1.csv | awk -F',' '{print $3}' | tail -n+2| sort | uniq -c | sort -nr | wc -l

Output:



Top 10 most frequent value:

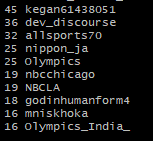
Code Explanation:

Open the csv file | split by ‘,’ print column 3 | unselect first line | count all unique values | sort rows | print first 10 lines

Code:

cat filtered\_tweets\_1.csv | awk -F',' '{print $3}' |tail -n+2 |sort | uniq -c | sort -nr | head -n10

Output:



1. Hashtags are commonly used in Twitter (e.g., #olympics2021). Write commands to count the number of hashtags in the “text” column of the data. Please notice that hashtags always start with #. If a hashtag appears twice in a text value, it should be counted twice.

Code Explanation:

Open the csv file | split by ‘,’ print column 2 | find and count character ‘#’

Code:

cat filtered\_tweets\_1.csv | awk -F',' '{print $2}' | grep -c '#'

Output:



1. In the following, let’s focus on certain columns contained in the file. Write commands to only keep columns that are in the following list:

* id
* user\_screen\_name
* user\_created\_at
* user\_followers
* user\_friends
* date

In particular, please only keep the tweets with user\_friends and user\_followers each larger than 1000. Export the selected data to a new file named “filtered\_tweets\_2.csv”. Please ensure that the file “filtered\_tweets\_2.csv” contains data from the selected columns as well as the column names.

Code Explanation:

Split the file by ‘,’ | if value in column 11 is greater than 1000 and value in column in 10 is greater than 1000, print column 1, 3 ,9, 10, 11, 12 and save in a new csv file called filtered \_tweets\_2.csv

Code:

awk -F',' '{if ($11 > 1000 && $10 > 1000)print $1","$3","$9","$10","$11","$12}' filtered\_tweets\_1.csv > filtered\_tweets\_2.csv

1. In the file “filtered\_tweets\_2.csv”, how many tweets are with a NA value in the last column (i.e., the column “date”)? How many user accounts were created prior to 2020. Please note that the column “user\_created\_at” specifies when a user account was created and one Twitter user might have produced multiple tweets and you are supposed to count the accounts.

**How many tweets with a NA I the date column:**

Code Explanation:

Open the csv file | split by ‘,’ if the column 6 is null or equal to “NA” print the line | count rows

Code:

cat filtered\_tweets\_2.csv | awk -F',' '{if ($6 == Null || $6 == "NA") print}' | wc -l

output:



**How many user accounts were created prior to 2020:**

Code Explanation:

Open the csv file | sort and remove duplicate id| split by ‘,’ print column 3 | split by ‘’ and print column 1 | split by ‘/’ select rows if column 3 is smaller than 2020 | remove first line | count total lines

Code:

cat filtered\_tweets\_2.csv | sort -u -t, -r -k1,1 | awk -F ',' '{print $3}' | awk -F ' ' '{print $1}' | awk -F '/' '{if ($3 < 2020) print}' | tail -n+2 | wc -l

output:

