Facebook Data Mining Assignment

Michael Edelman

May 2022

1 How many individual rows (in contrast to e.g., "lines") of data? By "row", I mean a self-contained unit of information in this data-set.

```
mikeede11@LAPTOP-F8JUS0BM:> csvstat facebookdata.csv | grep -i Row
Row count: 3222
mikeede11@LAPTOP-F8JUS0BM:>
```

2 List the columns names, in order.

```
mikeede11@LAPTOP-F8JUS0BM:> csvcut -n facebookdata.csv
  1: status_id
  2: status_message
  3: link_name
  4: status_type
  5: status_link
  6: status_published
  7: num_reactions
  8: num_comments
  9: num_shares
 10: num_likes
 11: num_loves
 12: num_wows
 13: num_hahas
 14: num_sads
 15: num_angrys
mikeede11@LAPTOP-F8JUS0BM:>
```

3 How many, of each status_type, are there? The query results must have n lines (one per status_type), sorted in increasing order of the number of a given status_type.

```
mikeede11@LAPTOP-F8JUS0BM:> csvsql --query "
SELECT COUNT(status_type) AS num_of_type, status_type
FROM facebookdata
GROUP BY status_type
ORDER BY num_of_type" facebookdata.csv |tr , " "
num_of_type status_type
481 photo
1337 video
1404 link
mikeede11@LAPTOP-F8JUS0BM:>
```

4 In this step, Extract (only) the contents of the status_message and all of the num_fields, displaying only the first three lines of the file. To make the implementation easier, your three lines of output should include the "csv header line".NOTE: I interpreted this as the header line counting as one of the three lines of output as opposed to the header line and 3 lines of row data output. in that case I would have used -n4.

```
mikeede11@LAPTOP-F8JUS0BM:> csvcut -c status_message,num_reactions,num_comments,num_shares,num_likes,num_loves,num_wows,num_hahas,num_sads,num_angrys \
>  facebookdata.csv | head -n3
status_message,num_reactions,num_comments,num_shares,num_likes,num_loves,num_wows,num_hahas,num_sads,num_angrys
"Ben Simmons will likely be the No. 1 pick of the NBA Draft, but who should it be?",5565,178,461,5488,43,13,19,0,2
"How to coach the ""Triangle Offense,"" as explained by Metta World Peace. (via QBronald/Twitter)",11997,1932,3158,10385,96,15,1499,0,2
mikeede11@LAPTOP-F8JUS0BM:>
```

5 In this step, Extract (only) the contents of the status_id and all of the num_fields, displaying the first three lines of data that begin at line 2 of the file. In other words, assume that lines are numbered 1..n; I don't want to see the "header line" so begin emitting lines beginning at the second line.

```
mikeede11@LAPTOP-F8JUS0BM:> csvcut -c status_id,num_reactions,num_comments,num_shares,num_likes,num_loves,num_wows,num_hahas,num_sads,num_angrys \
> facebookdata.csv | tail -n+2 | head -n3
7331091005_10154123560186006,5565,178,461,5488,43,13,19,0,2
7331091005_10154123362896006,11997,1932,3158,10385,96,15,1499,0,2
7331091005_10154123319126006,2063,270,400,1971,28,47,10,0,7
mikeede11@LAPTOP-F8JUS0BM:>
```

6 Determine the status_message with the most reactions (of all types, as defined above).

```
mikeede11@LAPTOP-F8JUS0BM:> csvsql --query "
> SELECT status_message, (num_reactions + num_comments + num_shares + num_likes + num_loves + num_wows + num_hahas + num_sads + num_angrys)
> AS total_popularity
> FROM facebookdata
> ORDER BY total_popularity DESC" facebookdata.csv | tail -n+2 | head -n1
LeBron and the Cavs are tired of being bullied,894597
mikeede11@LAPTOP-F8JUS0BM:>
```

7 fb.sh uploaded to git