Assignment 2 – Reading Computer Log files

Regular Expressions and Text Pattern Matching

STA141B Spring 2023

Professor Duncan Temple Lang

Due: Wednesday, May 3, 11pm

Submit via Canvas

As computers/operating systems run, different events occur that are logged and recorded in various "log" files. These include events such as

- attempted logins to the machine from other machines
- logins from the same machine
- running commands as another user (typically root) via the sudo command
- adding users

Monitoring attempted, successful and failed login attempts helps understand potential attacks on the machine. Understanding what is being done via sudo commands and by whom helps monitor changes to the system.

The MergedAuth.log file on Canvas is the single file you need to process and analyze for this assignment. It consists of 5 log files from different machines merged into a single file. The section for each of the 5 files starts with

```
# log-file-name
```

and is followed by the lines from the log file.

Each log-file message is a single line, i.e., no message spans multiple lines.

Consider 2 consecutive lines

```
Nov 30 15:22:39 ip-172-31-27-153 sshd[22844]: Invalid user admin from 124.205.250.51
Nov 30 15:22:39 ip-172-31-27-153 sshd[22844]: input_userauth_request: invalid user admin [preauth]
```

These come from the log file auth.log identified by

```
# auth.log
```

The structure of the first and **most** other lines is as follows:

```
Nov 30 15:22:39 ip-172-31-27-153 sshd[22844]: Invalid user admin from 124.205.250.51
|_____________| |_____________| |__| |___|   |____________________________________|
   date-time      logging host   app   PID                  message

<----------------- METADATA ----------------->
```

Your first task is to create a data.frame with a row for each actual log message line, with the variables

- date-time
- the name of the host collecting the log messages
- the app(lication)
- the process ID (PID)
- the message, i.e, the remainder of the log line
- the name of the log file from which the line came, e.g. auth.log, SSH_2k.log

If there is no app or PID for a given line, set the corresponding value to `NA`.

Data Validation and Exploration

- Verify that the PIDs are all numbers.
- How many lines are in each log file?
- What are the range of date-times for the messages? for each of the different log files in the combined file? How many days does each log file span?
- Do the application names contain numbers? If so, are they just versions, e.g. ssh2, or is there additional structure to the numbers?

- Is the host value constant/the same for all records in each log file?
- What are the most common apps (daemons/programs) that are logging information on each of the different hosts?

Logins - valid and invalid

There are messages such as

```
Mar 27 13:08:09 ip-10-77-20-248 sshd[1361]: Accepted publickey for ubuntu from 85.245.107.41 port 54259 ssh2 .
Dec 31 22:27:48 ip-172-31-27-153 sshd[8003]: Invalid user admin from 218.2.0.133
Mar 28 20:21:20 ip-10-77-20-248 systemd-logind[1118]: New session 50 of user ubuntu.
Mar 28 20:21:52 ip-10-77-20-248 sshd[30039]: Connection from 85.245.107.41 port 63502 on 10.77.20.248 port 222
Mar 28 22:36:12 ip-10-77-20-248 sshd[30174]: error: maximum authentication attempts exceeded for invalid user
```

- Find valid/successful logins
  - What are the user names and Internet Protocol (IP) addresses of those logins
- Find the invalid user login/ids
  - What were the associated IPs
  - Were there multiple invalid user logins from the same IP addresses
  - Were there valid logins from those IP addresses
  - Are there multiple IPs using the same invalid login?
    - Are these related IPs, e.g., from the same network/domain?
- What IP address had too many authentication failures.

Sudo commands

- What are the executables/programs run via sudo
  - By what user
  - What machine are they on?

Useful Functins

- grep(), grepl()
- gregexpr(), regexpr()
- gsub(), sub()
- readLines()
- strsplit(), substring()
- paste(), paste0(), sprintf()
- by(), tapply(), split()
- cumsum()
- strptime(), as.POSIXct()
- trimws()
- strptime(), as.POSIXct(), difftime()

Report and Submission

Submit

- a PDF document for your report
- any R files containing function definitions that are used in report (and not defined directly in the report)
- your .Rmd file if you used Rmarkdown.

The report should answer the questions posed above and also describe the strategy for solving them. The narrative should persuade me and others that your approach and results are correct, i.e., that you did the right things and did it correctly.

Resources

- Mastering Regular Expressions, Jeffrey Friedl, O'Reilly
- https://www.regular-expressions.info
- https://www.debuggex.com
- R help pages.