**CIT 383 – Scripting 1**

**Musgrave**

**Spring 23**

**Lab 8**

**Tasks**

Develop a script that parses Apache Common Log Format log files and produces a report. The data for each of the two reports is to be stored in one of two counting dictionaries, which are then used to generate the output for the two reports.

The default layout for an Apache access log is defined by httpd.conf directives as follows:

LogFormat "%h %l %u %t \"%r\" %>s %b \"%{Referer}i\" \"%{User-Agent}i\""

where:

* %h is the remote host (i.e., the client IP)
* %l is the identity of the user determined by identd (not usually used since not reliable)
* %u is the user name determined by HTTP authentication (not used in the sample)
* %t is the time the server finished processing the request including the GMT offset
* %r is the full HTTP request line from the client. ("GET / HTTP/1.0")
* %>s is the status code sent from the server to the client (200, 404 etc.)
* %b is the size of the response to the client (in bytes).
* %{Referer} is the referring page (e.g., the home page that has a hyperlink)
* %{UserAgent} is the browser itself, acting on behalf of the user

An example log file apache-access\_log is provided on canvas which you should use for testing purposes.

A typical line in the sample file will look like this:

192.168.1.142 - - [30/Sep/2019:16:19:18 -0400] "GET / HTTP/1.1" 403 4897 "-" "Mozilla/5.0 (Windows NT 10.0; WOW64; rv:49.0) Gecko/20100101 Firefox/49.0"

The dashes are used as placeholders for unpopulated fields. To parse such a line into fields, we first need to read the file to get the lines.

Your script must generate the following report with two specific sub-reports. The output should follow this format. You may use the output below to test that your script is counting correctly

1. Start with a descriptive header followed by at least one blank line, using the variable for the file name:

----------------------------------------------------

Statistics for the Apache log file access\_log

-----------------------------------------------------

2. The first sub-report is a histogram of IP addresses (in a field of 20 spaces) using the asterisk (\*) character, with a short header to explain what the data represent. For example:

Frequency of Client IP Addresses:

::1 \*\*\*\*\*\*\*

192.168.1.142 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

192.168.1.1 \*\*\*\*\*\*\*\*\*\*\*\*\*

192.168.1.34 \*\*\*\*

192.168.1.127 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

192.168.1.56 \*\*\*\*\*\*\*\*\*

192.168.1.138 \*\*\*\*\*\*\*\*\*

192.168.1.156 \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

3. The second sub-report is a list of HTTP status codes with percentages of URLs that resulted in each status code, which should be displayed as shown. The percentage is based on the count of each unique code divided by the total number of lines in the file, which you should derive from the size of the List of lines.

HTTP Status Codes Summary:

200: 45.37%

403: 11.11%

404: 35.19%

301: 5.56%

304: 2.78%

In solving the problem please implement the following functions:

* getFileData(): This function takes the name of the log file as argument and returns list containing all the data in the file specified
* genClientIP(): This function takes the content of the log file(data return by the getFileData function) as argument and returns a dictionary with the client IP addresses as keys and frequency of IP addresses as values
* genStatusCode():This function takes the content of the log file(data return by the getFileData function) as argument and returns a dictionary with the Status Codes as keys and frequency of Status codes as values
* printReport():This function takes as arguments, the IP Address and Status Code dictionaries and generates the reports as indicated above. The output of your report should be properly formatted. Please note that, in printing the Status code summary report, the frequency of occurrence of the status codes have to be converted to percentages.

Note: Use regular expressions to extract IP addresses and Status codes.