Bright Network IEUK Project Task: Engineering

James Omorege John

I used Python to look at a sample web server log file for this task. I used collections and re (regular expressions).Counter libraries to find and count the most common IP addresses, URLs, and User-Agent strings. This helped me figure out which users, pages, and devices were using the server the most.

I began by making a regular expression pattern that would match various parts of the log, like the IP address, date and time, request method, URL, status code, referrer, and user agent. I used the re.match() method to get these out, and then I used Counter() to keep track of how many times each unique value appeared.

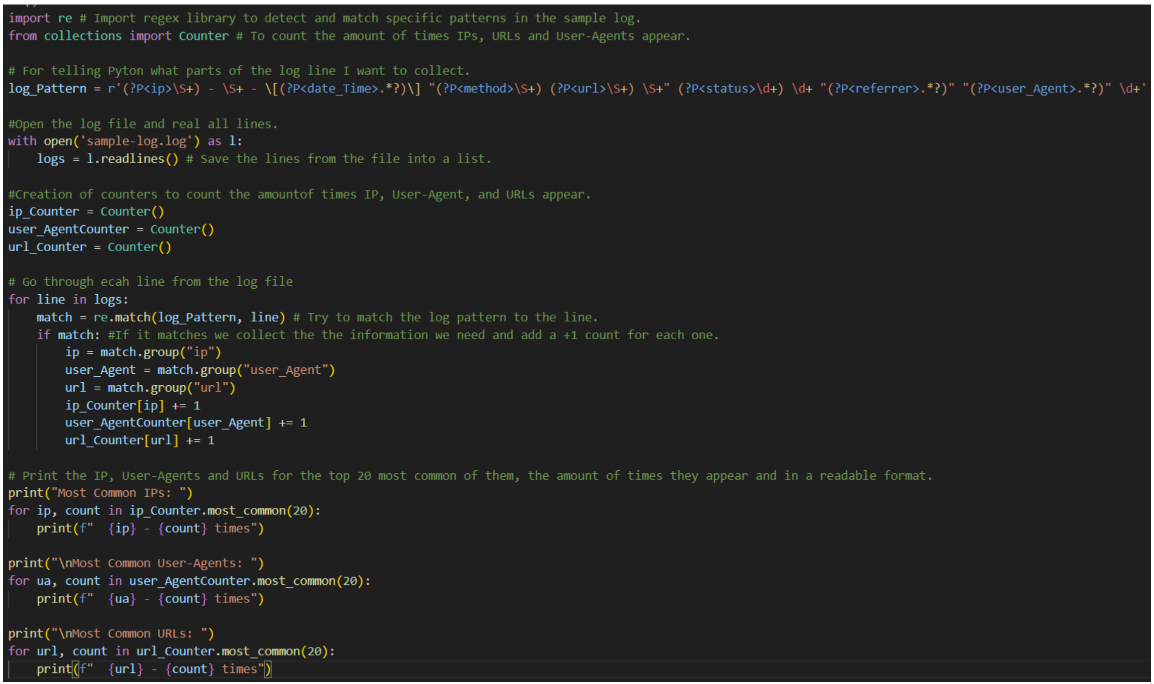
Based on what I found, the top IP addresses showed up thousands of times. The top two IPs, 45.133.1.1 and 45.133.1.2, each showed up 5,400 times. This could mean that bots are visiting your site or that automated access is happening repeatedly.

A Windows Chrome browser was the most common User-Agent, with 44,882 instances. I did however see a lot of automation tools, such as curl, wget, sqlmap, and even Burp Suite. This strongly suggests that some of the traffic may be bots or tools for penetration testing, not real users.

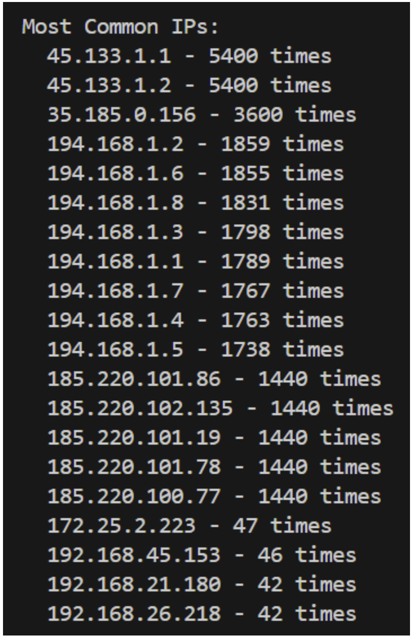
The pages that got the most hits were /episodes/ep-42-synthesizer-history, /contact, and /about, each of which got more than 15,000 hits. These are probably pages that are open to the public or that a lot of people visit.

This analysis helped me get hands-on experience with log parsing and spotting suspicious activity like bot traffic. If this were a real-world scenario, I’d recommend further investigation into the IPs and agents involved, and potentially implementing bot protection or rate limiting to reduce unnecessary or harmful traffic.

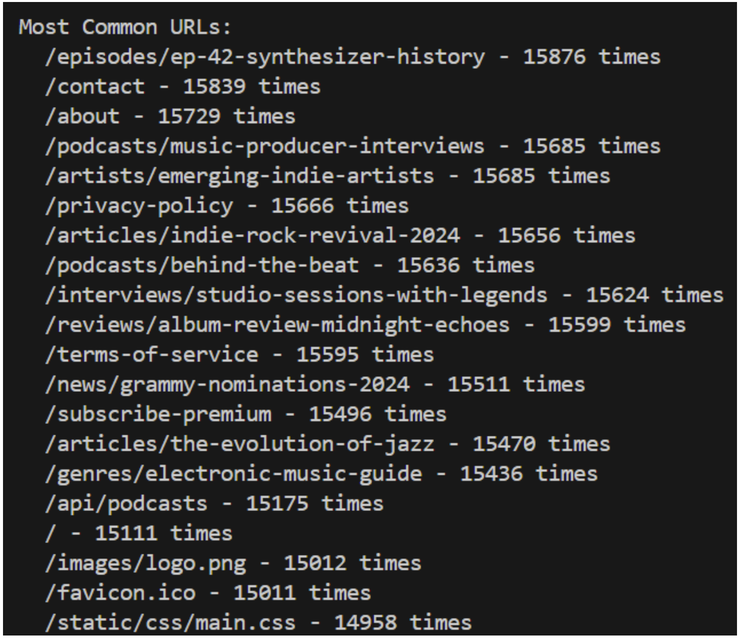
Pictures of my results and code are found below.



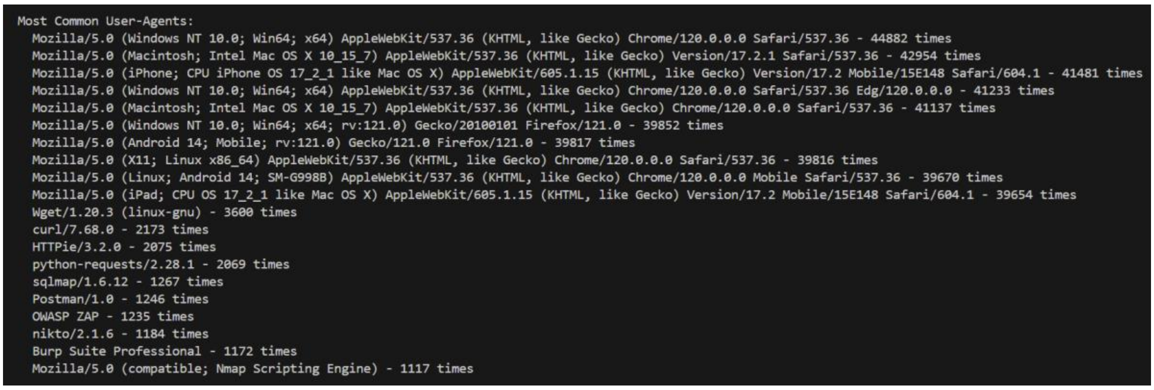
```python
import re # Import regex library to detect and match specific patterns in the sample log.
from collections import Counter # To count the amount of times IPs, URLs and User-Agents appear.

# For telling Pyton what parts of the log line I want to collect.
log_Pattern = r'(?P<ip>\S+) - \S+ - \[(?P<date_Time>.*?)\] "(?P<method>\S+) (?P<url>\S+) \S+" (?P<status>\d+) \d+ "(?P<referrer>.*?)" "(?P<user_Agent>.*?)" \d+'

#Open the log file and real all lines.
with open('sample-log.log') as l:
    logs = l.readlines() # Save the lines from the file into a list.

#Creation of counters to count the amountof times IP, User-Agent, and URLs appear.
ip_Counter = Counter()
user_AgentCounter = Counter()
url_Counter = Counter()

# Go through ecah line from the log file
for line in logs:
    match = re.match(log_Pattern, line) # Try to match the log pattern to the line.
    if match: #If it matches we collect the the information we need and add a +1 count for each one.
        ip = match.group("ip")
        user_Agent = match.group("user_Agent")
        url = match.group("url")
        ip_Counter[ip] += 1
        user_AgentCounter[user_Agent] += 1
        url_Counter[url] += 1

# Print the IP, User-Agents and URLs for the top 20 most common of them, the amount of times they appear and in a readable format.
print("Most Common IPs: ")
for ip, count in ip_Counter.most_common(20):
    print(f"  {ip} - {count} times")

print("\nMost Common User-Agents: ")
for ua, count in user_AgentCounter.most_common(20):
    print(f"  {ua} - {count} times")

print("\nMost Common URLs: ")
for url, count in url_Counter.most_common(20):
    print(f"  {url} - {count} times")
```



```
Most Common IPs:
  45.133.1.1 - 5400 times
  45.133.1.2 - 5400 times
  35.185.0.156 - 3600 times
  194.168.1.2 - 1859 times
  194.168.1.6 - 1855 times
  194.168.1.8 - 1831 times
  194.168.1.3 - 1798 times
  194.168.1.1 - 1789 times
  194.168.1.7 - 1767 times
  194.168.1.4 - 1763 times
  194.168.1.5 - 1738 times
  185.220.101.86 - 1440 times
  185.220.102.135 - 1440 times
  185.220.101.19 - 1440 times
  185.220.101.78 - 1440 times
  185.220.100.77 - 1440 times
  172.25.2.223 - 47 times
  192.168.45.153 - 46 times
  192.168.21.180 - 42 times
  192.168.26.218 - 42 times
```



```
Most Common URLs:
  /episodes/ep-42-synthesizer-history - 15876 times
  /contact - 15839 times
  /about - 15729 times
  /podcasts/music-producer-interviews - 15685 times
  /artists/emerging-indie-artists - 15685 times
  /privacy-policy - 15666 times
  /articles/indie-rock-revival-2024 - 15656 times
  /podcasts/behind-the-beat - 15636 times
  /interviews/studio-sessions-with-legends - 15624 times
  /reviews/album-review-midnight-echoes - 15599 times
  /terms-of-service - 15595 times
  /news/grammy-nominations-2024 - 15511 times
  /subscribe-premium - 15496 times
  /articles/the-evolution-of-jazz - 15470 times
  /genres/electronic-music-guide - 15436 times
  /api/podcasts - 15175 times
  / - 15111 times
  /images/logo.png - 15012 times
  /favicon.ico - 15011 times
  /static/css/main.css - 14958 times
```



```
Most Common User-Agents:
  Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/120.0.0.0 Safari/537.36 - 44882 times
  Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Version/17.2.1 Safari/537.36 - 42954 times
  Mozilla/5.0 (iPhone; CPU iPhone OS 17_2_1 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/17.2 Mobile/15E148 Safari/604.1 - 41481 times
  Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/120.0.0.0 Safari/537.36 Edg/120.0.0.0 - 41233 times
  Mozilla/5.0 (Macintosh; Intel Mac OS X 10_15_7) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/120.0.0.0 Safari/537.36 - 41137 times
  Mozilla/5.0 (Windows NT 10.0; Win64; x64; rv:121.0) Gecko/20100101 Firefox/121.0 - 39852 times
  Mozilla/5.0 (Android 14; Mobile; rv:121.0) Gecko/121.0 Firefox/121.0 - 39817 times
  Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/120.0.0.0 Safari/537.36 - 39816 times
  Mozilla/5.0 (Linux; Android 14; SM-G998B) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/120.0.0.0 Mobile Safari/537.36 - 39670 times
  Mozilla/5.0 (iPad; CPU OS 17_2_1 like Mac OS X) AppleWebKit/605.1.15 (KHTML, like Gecko) Version/17.2 Mobile/15E148 Safari/604.1 - 39654 times
  Wget/1.20.3 (linux-gnu) - 3600 times
  curl/7.68.0 - 2173 times
  HTTPie/3.2.0 - 2075 times
  python-requests/2.28.1 - 2069 times
  sqlmap/1.6.12 - 1267 times
  Postman/1.0 - 1246 times
  OWASP ZAP - 1235 times
  nikto/2.1.6 - 1184 times
  Burp Suite Professional - 1172 times
  Mozilla/5.0 (compatible; Nmap Scripting Engine) - 1117 times
```