prompt.adoc 2.3 KB

Exercise - Parse HTML

Learning Objectives

- CNNI002 - Employ commands using common shells
  - CNNI002.003 - Demonstrate appropriate use of pipes and redirection
  - CNNI002.005 - Identify methods of gaining more information about commands and switches
- CCNI003 - Analyze the Linux file system
  - CCNI003.007 - Describe regular expressions
  - CCNI003.008 - Create regular expressions to find data within in the file system
  - CCNI003.009 - Identify the information that a regular expression will return

Learning Outcomes

- Familiarity with regular expressions, their syntax and their application
- Confidence generating simple regular expressions to match specific patterns
- Familiarity with Extended and Perl grep (`grep -{E,P}`)

Scenario

You have pulled back a webpage to a Linux system and need to parse out specific information for analysis.
Your parsing mechanism needs to be automatic and portable to other webpages.
Be sure to submit the actual counts, not the md5sum's, along with the command(s) you used.

Activity

- Download `/usr/share/linux/page.html` (Verify your file - MD5SUM: `fc913eefe9aae1d4a26f3ddd4a6d4134`)
- Identify the following flags by parsing the webpage using regular expressions and other linux binaries

Check Your Work

```
Check Your Work:
$ echo <#> | md5sum

FLAG 1:    Count total # of times "<link" tag appears at the begining of a line in the html file:     c30f7472766d25af1dc80b3ffc9a58c7
FLAG 2:    Count total # of times "OWASP" appears in the html file.                                   51aa762ea14b40fb8876c887eea6adb6
FLAG 3:    Count total # of times "<div" appears anywhere in the html file.                           8faff61bc1198cc6bdc19adafc27fc82
FLAG 4:    Count total # of times "<div" appears at the begining of a line in the html file.          367764329430db34be92fd14a7a770ee
FLAG 5:    Count total # of times "</script>" appears anywhere in the html file.                      367764329430db34be92fd14a7a770ee
FLAG 6:    Count total # of times "</script>" appears at the end of a line in the html file.          7c5aba41f53293b712fd86d08ed5b36e
FLAG 7:    Count total # of times "<head>" appears anywhere in the html file.                         b026324c6904b2a9cb4b88d6d61c81d1
FLAG 8:    Count total # of URLs (that start with 'http://' or 'https://') in the html file.          7162e6c85b5430f81c374d194909ad60
FLAG 9:    Count total # of unique URLs (that start with 'http://' or 'https://') in the html file.   091b847d9660c0555bc6f29994a03710
FLAG 10:   Count total # of emails in the html file.                                                  21d192a7f99b9d042621777150ea8a5f
FLAG 11:   Count total # of unique emails in the html file.                                           12fc54257f1ae2301d3f9db906d7f414
FLAG 12:   Count total # of phone numbers in the html file.                                           b026324c6904b2a9cb4b88d6d61c81d1
FLAG 13:   Count total # of CWEs in the html file.                                                    7c67493bd72ceff21059c3d924d17518
FLAG 14:   Count total # of unique CWEs in the html file.                                             2a53da1a6fbfc0bafdd96b0a2ea29515
```

Deliverables

- Libre Office Document containing count and command used to obtain the count, example:

Flag	Count	Command
1	45	grep -Po "([0-9]{1,3}\.){3}[0-9]{1,3}" page.html | wc -l

Grading

50% - Flags 1-7 collected
10% - Flag 8 collected
10% - Flag 9 collected
10% - Flag 10 collected
5% - Flag 11 collected
5% - Flag 12 collected
5% - Flag 13 collected
5% - Flag 14 collected

Useful Resources

- E. Gamet Grep Quick Reference Chart [Online] Available: 📎 http://www.ericagamet.com/wp-content/uploads/2016/04/Erica-Gamets-GREP-Cheat-Sheet.pdf (http://www.ericagamet.com/wp-content/uploads/2016/04/Erica-Gamets-GREP-Cheat-Sheet.pdf)
- **RFC 3986** (URI/URL): Generic Syntax Ch. 2)), **RFC 3696** (Restrictions on email addresses Ch. 3)
- Regexer: http://www.regexr.com (http://www.regexr.com)