Log File Analysis using RegEx - Brief ReadMe

Tahmeed Tureen

Fall 2017

This document briefly highlights what I did for this small scale project regarding filtration of a web-server access log file to analyze the primary sources of hacking attempts. The access log file was provided by Dr. Chris Teplovs at the School of Information, University of Michigan.

Steps:
1. Parse & filter the log file using regular expressions in Python
2. Filter the rows of the log into two output files: (i) valid, & (ii) invalid
3. Definition of valid is

(Courtesy of Dr. Teplovs):
(a) The HTTP verb is GET or POST
(b) AND the status code is 200
(c) AND the URL being accessed starts with http:// or https://, followed by one or more alphabetic characters (i.e. not a digit or a symbol). For example, the URL should NOT start with 'http:///', which is an error.
(d) AND the top-level domain consists of only letters. This is to say, if the host name is actually
a numerical IP address like '202.96.254.200', we dont count it. If the whole domain name is just '.com' as in http://.com/blah or does not even contain a dot as in http://c/blah, we do not count it

Thank you for reading