Python Programming

Task 1

Tomas Raila

February 16, 2020

Task description

- **The task**: write an HTTP log parser in Python.
- The parser should work with Common Log Format files: https://en.wikipedia.org/wiki/Common_Log_Format
- Sample log file: https://raw.githubusercontent.com/elastic/examples/master/CommonDataFormats/apache_logs/apache_logs
- You can also look for other similar log files (but make sure the format is the same): https://www.google.co.uk/search?q=inurl:access.log+filetype:log
- **Note**: most of such log files might contain two extra attributes at the end of each line: request URL and browser identification string. Your parser should ignore them.

Requirements

The parser must be able to:

1. Group the logged requests by IP address or HTTP status code (selected by user).
2. Calculate one of the following (selected by user) for each group:
   1. Request count
   2. Request count percentage of all logged requests
   3. Total number of bytes transferred
3. Print the results in descending order.
   Note: order the results by values described in (2), not by IP address or HTTP code.
4. Optionally limit the number of rows printed (specified by user)

All parameters (including the input file name) should be passed to the parser script from command line. Parameters for (1) and (2) are required, parameter for (4) is optional.

Other info

Maximum grade for the task: **1.0**

Grading criteria:

- The parser should work (correctly) with any log file of CLF format.
- The code should be clean and pythonic. Please do not write Java or C code in Python!
- The parser should handle unexpected situations (i.e. empty log file, incorrectly specified command line arguments, etc.)

The deadline for task submission to Gitlab is: **2020-03-09 10:00** (beginning of 1st exercise session).

The code also has to be presented and explained to lecturer in class.

See task submission instructions in Lecture 1 slides.