Manipulating Strings

Gaston Sanchez

September 26, 2016

Strings Manipulation

We'll be working with the data file `may-logs.txt` available in the course's github repository. This file is a log file created by a server of a given web site. Basically, `may-logs.txt` is a text file in *common log format*:

https://en.wikipedia.org/wiki/Common_Log_Format

http://www.jafsoft.com/searchengines/log_sample.html

One of the fields stored in a log file is the HTTP status code:

https://en.wikipedia.org/wiki/List_of_HTTP_status_codes

Questions

Look at the log file `may-logs.txt`. Assume that an IP address is used for a unique user.

- How would you read the data in R?
- How many lines contain `.jpg` files? `.png` files? `.gif` files? `.ico` files?
- How many lines contain image files?
- How many **unique** `.jpg` files? unique `.png` files? unique `.gif` files? unique `.ico` files?
- How many unique image files?
- How many success status (200's)?
- How many redirection status (300's)?
- How many client error status (400's)?
- How many server error status (500's)?
- What is the most common client error code?
- Obtain "visits" = IP address + date(dd/MM/yyyy)
- What is the day of the month with more visits?

```
# number of lines with jpg, png, gif, ico files
grep -o "/.*\.jpg" may-logs.txt | wc -l
grep -o "/.*\.png" may-logs.txt | wc -l
grep -o "/.*\.gif" may-logs.txt | wc -l
grep -o "/.*\.ico" may-logs.txt | wc -l


# number of lines with image files
# (matching just the file extension)
grep -o "\.[jpgi][pnic][ggfo]" may-logs.txt | wc -l


# number of unique jpg, png, gif, ico files
grep -o "/.*\.jpg" may-logs.txt | sort | uniq | wc -l
grep -o "/.*\.png" may-logs.txt | sort | uniq | wc -l
grep -o "/.*\.gif" may-logs.txt | sort | uniq | wc -l
grep -o "/.*\.ico" may-logs.txt | sort | uniq | wc -l


# how many unique image files?
# (extended grep to get just the name of the image file)
grep -oE "\\w+\.png" may-logs.txt | sort | uniq
grep -oE "\\w+\.jpg" may-logs.txt | sort | uniq | wc -l
grep -oE "\\w+\.png" may-logs.txt | sort | uniq | wc -l
grep -oE "\\w+\.gif" may-logs.txt | sort | uniq | wc -l
grep -oE "\\w+\.ico" may-logs.txt | sort | uniq | wc -l


# number of unique image files
grep -oE "\\w+\.[jpgi][pnic][ggfo]" may-logs.txt | sort | uniq | wc -l
grep -oE "\\w+\.[jpgi][pnic][ggfo]" may-logs.txt > images.txt


# define a "visit" as a combinatino of IP address with date
cut -f 1 -d ":" may-logs.txt | head -n 5

# day with most "visits"
cut -f 1 -d ":" may-logs.txt | sort | uniq | cut -f 2 -d "[" | sort | uniq -c | sort
```