Traffic report

The hackermen from the previous exercise are not naive. They know that corporations are constantly monitoring what they do looking for slightest signs of insubordination. But hackermen still have an advantage — they managed to access all network logs, so they can analyze the corporations' steps and prepare an attack. Every line in logs has the following format:

<IPv4 address> [<datetime>] "<HTTP request>" <HTTP response code> <bytes sent>

e.g.:

10.4.180.222 [28/Jan/2018:13:52:04 +0100] "GET http://www.google.com/ HTTP/1.1" 200 2326

The first step of the analysis will be checking which URLs, and how frequently, are requested by corporations,. That's your job!

Requirements

Write a script page_report.py that generates a report from a log file. Your script should count requests for each URL, ignoring the protocol, ending slash and query string parameters — they all should be stripped.

The log file path should be read from the command line argument and the generated report written to the standard output. Every line of the result should follow the CSV format:

"<stripped url>",<requests count>

e.g.:

"www.google.com",1

The records in the result should be sorted by the number of requests in descending order, and if two URLs are requested equally often, they should be sorted lexicographically.

Extra points will be given for:

- good handling of very big files
- input validation (invalid lines should be ignored and the number of such lines should be printed on stderr, e.g. "Invalid log lines: 42")

Example

Given the following log file:

today.log

```
10.4.180.222 [28/Jan/2018:10:02:32 +0100] "GET http://clearcode.cc/ HTTP/1.1" 200 1080
10.4.180.222 [28/Jan/2018:10:03:31 +0100] "GET http://www.clearcode.cc HTTP/1.1" 200 3056
10.4.180.222 [28/Jan/2018:10:05:30 +0100] "GET http://clearcode.cc/careers HTTP/1.1" 200 3056
10.4.180.222 [28/Jan/2018:10:08:29 +0100] "GET http://clearcode.cc/careers/ HTTP/1.1" 200 3056
10.4.180.222 [28/Jan/2018:10:13:29 +0100] "GET http://clearcode.cc/careers? HTTP/1.1" 200 3056
10.4.180.222 [28/Jan/2018:10:21:27 +0100] "GET http://clearcode.cc/careers/? HTTP/1.1" 200 3056
10.4.180.222 [28/Jan/2018:10:34:26 +0100] "GET
http://clearcode.cc/careers?offer=internship&type=python HTTP/1.1" 200 4545
```

```
10.4.180.222 [28/Jan/2018:10:55:25 +0100] "GET
http://clearcode.cc/careers?type=frontend&offer=internship HTTP/1.1" 200 5454
```

and running the following command:

```
python page_report.py today.log > report.csv
```

we should get the following report:

report.csv

```
"clearcode.cc/careers",6
"clearcode.cc",1
"www.clearcode.cc",1
```