Python for Practice – Homework 3

The goal of this homework is to implement several queries over the web-access dataset. (The homework corresponds to exercises e21 and e22.)

Query #1

Load the file web-access/access_log as a dataframe. Parse it with a regular expression to extract the address and date. Print out a few lines using method show(truncate=False).

Sample output (note that your results or their actual representation may differ):

```
+--------------------------+-------------------------+
|addr                      |date                     |
+--------------------------+-------------------------+
|in24.inetnebr.com         |01/Aug/1995:00:00:01 -0400|
|uplherc.upl.com           |01/Aug/1995:00:00:07 -0400|
|uplherc.upl.com           |01/Aug/1995:00:00:08 -0400|
|uplherc.upl.com           |01/Aug/1995:00:00:08 -0400|
|uplherc.upl.com           |01/Aug/1995:00:00:08 -0400|
|ix-esc-ca2-07.ix.netcom.com|01/Aug/1995:00:00:09 -0400|
|uplherc.upl.com           |01/Aug/1995:00:00:10 -0400|
|slppp6.intermind.net      |01/Aug/1995:00:00:10 -0400|
|piweba4y.prodigy.com      |01/Aug/1995:00:00:10 -0400|
|slppp6.intermind.net      |01/Aug/1995:00:00:11 -0400|
|slppp6.intermind.net      |01/Aug/1995:00:00:12 -0400|
|ix-esc-ca2-07.ix.netcom.com|01/Aug/1995:00:00:12 -0400|
|slppp6.intermind.net      |01/Aug/1995:00:00:13 -0400|
|uplherc.upl.com           |01/Aug/1995:00:00:14 -0400|
|133.43.96.45              |01/Aug/1995:00:00:16 -0400|
|kgtyk4.kj.yamagata-u.ac.jp|01/Aug/1995:00:00:17 -0400|
|kgtyk4.kj.yamagata-u.ac.jp|01/Aug/1995:00:00:18 -0400|
|d0ucr6.fnal.gov           |01/Aug/1995:00:00:19 -0400|
|ix-esc-ca2-07.ix.netcom.com|01/Aug/1995:00:00:19 -0400|
|d0ucr6.fnal.gov           |01/Aug/1995:00:00:20 -0400|
+--------------------------+-------------------------+
```

Query #2

Print out ten visitors with most hits, ten urls with most hits and ten html pages (i.e. urls ending with .html) with most hits. Save the each of those top ten to a corresponding CSV file.

Sample output (note that your results or their actual representation may differ):

```
+-------------------+-----+
|addr               |count|
+-------------------+-----+
|edams.ksc.nasa.gov |6530 |
|piweba4y.prodigy.com|4846 |
|163.206.89.4       |4791 |
|piweba5y.prodigy.com|4607 |
|piweba3y.prodigy.com|4416 |
|www-d1.proxy.aol.com|3889 |
|www-b2.proxy.aol.com|3534 |
|www-b3.proxy.aol.com|3463 |
|www-c5.proxy.aol.com|3423 |
|www-b5.proxy.aol.com|3411 |
+-------------------+-----+

+-----------------------------------+-----+
|url                                |count|
+-----------------------------------+-----+
|/images/NASA-logosmall.gif         |97293|
|/images/KSC-logosmall.gif          |75283|
```

/images/MOSAIC-logosmall.gif	67356
/images/USA-logosmall.gif	66975
/images/WORLD-logosmall.gif	66351
/images/ksclogo-medium.gif	62670
/ksc.html	43619
/history/apollo/images/apollo-logo1.gif	37806
/images/launch-logo.gif	35119
/	30123

url	count
/ksc.html	43619
/shuttle/missions/sts-69/mission-sts-69.html	24592
/shuttle/missions/missions.html	22429
/software/winvn/winvn.html	10343
/history/history.html	10111
/history/apollo/apollo.html	8973
/shuttle/countdown/liftoff.html	7858
/history/apollo/apollo-13/apollo-13.html	7160
/shuttle/technology/sts-newsref/stsref-toc.html	6506
/shuttle/missions/sts-69/images/images.html	5261