Assignment #1

Course: ISA 414

Instructor: Dr. Arthur Carvalho

Points:100

Due date: September 16th, 2022, before 11:59 pm

Submission instructions: this assignment is to be done individually. All your answers should be inside a single *.ipynb* file. Your code must be well formulated (*i.e.*, no errors) and sound (*i.e.*, it does what the question asks it to do). In particular, one must be able to open your notebook file using VS Code and run the code without running into errors. Code with errors may receive zero points. Submit the final document on Canvas before the due date.

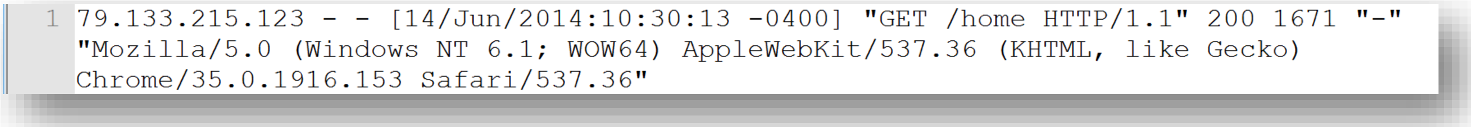
Background story: Don Draper, the CMO of the company you work for, loved the data-driven solution you proposed to solve the question: *which countries should one expect to have a high demand for butler services*? Recall that your solution consists of deriving user locations from IP addresses, which in turn are inside web logs generated by Apache web servers (see Lecture 4 and 5).

The idea now is to fully automate the whole analytics process. In particular: 1) the raw web log data will be automatically saved in a Hadoop environment at the end of each day; 2) once a month, a Python script will preprocess the web logs by imposing a tabular structure on them; and 3) users' locations will be derived from their IP addresses and sent straight to Don Draper's computer, who will visualize the results using an application called Tableau.

Mr. Anderson, also known as Neo, was the person responsible for coding the script required by step 2. Unfortunately, it seems that Mr. Anderson disappeared after taking some red pill.[1] Now it is up to you to do his job.

[1] This is a not-so-funny geek joke related to the movie The Matrix.

Question 1: by the end of this question, each log line in the file *access.log.txt* must be stored as one observation (row) in a CSV file. For example, the following log line



```
79.133.215.123 - - [14/Jun/2014:10:30:13 -0400] "GET /home HTTP/1.1" 200 1671 "-"
"Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko)
Chrome/35.0.1916.153 Safari/537.36"
```

becomes the following observation (row):

IP	Date	Request	Page	HTTP_Version	First_Code	Second_Code	User_Agent
79.133.215.123	14/Jun/2014:10:30:13 -0400	GET	/home	HTTP/1.1	200	1671	Mozilla/5.0 (Windows NT 6.1; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/35.0.1916.153 Safari/537.36

You should start by importing the file "*access.log.txt*" (available on Canvas) to Python using the command:

```
file = open('access.log.txt', 'r')

web_logs = file.readlines()
```

The variable web_logs is a list from a real-life Apache web server containing 360,000 elements, where each element represents a single log line.

a) Extract the *IP* variable from web_logs. The result should be a single list of length 360,000, where each element contains a single IP address. **[10 points]**

b) Extract the variable *Date* from web_logs. The result should be a single list of length 360,000, where each element contains a single date without the enclosing brackets, but including the time zone. **[10 points]**

c) Extract the variable *Request* from web_logs. The result should be a single list of length 360,000, where each element contains a single request (*e.g.*, GET) *without* quotation characters. **[10 points]**

d) Extract the variable *Page* from web_logs. The result should be a single list of length 360,000, where each element contains the path of a page, which starts at the character / (*e.g.*, /home). **[10 points]**

e) Extract the variable *HTTP_Version* from web_logs. The result should be a single list of length 360,000, where each element contains a numeric value describing the network protocol and version (*e.g.*, HTTP/1.1). **[10 points]**

f) Extract the variable *First_Code* from web_logs. The result should be a single list of length 360,000, where each element contains a numeric value (*e.g.*, 200). **[10 points]**

g) Extract the variable *Second_Code* from web_logs. The result should be a single list of length 360,000, where each element contains a numeric value (*e.g.*, 1671). **[10 points]**

h) Extract the variable *User_Agent* from web_logs. The result should be a single list of length 360,000, where each element contains a single user agent *without* quotation marks. **[10 points]**

i) Create a data frame by combining all the lists created in a) to h). **[10 points]**

j) Save the data frame created above to a CSV file called *data.csv*. Make sure the CSV file does not contain row names (index). **[5 points]**

k) Cleanliness: your code should not have any unnecessary/loose statements, such as irrelevant python cells. Failure to comply with this requirement means an automatic 5-point deduction (no partial credit) **[5 points]**