Python for webscraping module
Syllabus

Gabriele Rovigatti
email: gabriele.rovigatti@gmail.com

I Syllabus

Program - overview The aim of the course is both to provide the students with the basics of Pyhton (3.X.X) and to let them start working on simple webscraping tasks. At the end of the lessons, they are expected to be able to scrape standard websites, extract usable information and store it in machine-readable format - i.e., to obtain datasets for statistical analyses starting from plain webpages.

The lectures are introductory and are meant to provide the audience with the description of:

- Pyhton basic syntax;
- Python key strengths for web scraping;
- the most widely used tools and modules;
- programming best practices (secondary).

The lessons are structured to be as interactive as possible, therefore everyone is strongly encouraged to attend with her/his computer. I will upload course slide and the relative pieces of code directly into the course Dropbox folder. You can find the presentation and the codes of a previous special lecture here. All the codes, in particular, are in .zip format (direct download here): you can download it and extract all files in the same directory. Windows

users may then call the functions directly from hyperlinks in the presentation, Mac and Linux users will find the relative pieces of code within the "/code/" directory.

Course prerequisites Below you can find the (few) requirements to attend the lectures:

- Throughout the lecture I will present many pieces of code. Most of them are general, however they have been written and tested on Python **3.6**. I suggest to download and install its last version (available here for any OS). Make sure to install `pip` (or any other module installer available) and the Shell. Lastly, the code for previous lectures had been tested on Python 2.7, hence may not work properly on 3.6.
- We will use (not so many) modules: these must be downloaded and installed by `pip install` or `easy_install` before calling them. In particular, we will use: `requests`, `selenium`, `time`, `bs4` and `csv`. Please make sure to have them installed before the class start. In case of doubts, follow the instructions provided here and here
- A basic knowledge of the concepts of function, for loop, lists and/or arrays is assumed.

Resources: Below some useful resources available online.

- Detailed guide for webscraping and data analysis with `BeautifulSoup` (with Python 3) - https://www.dataquest.io/blog/web-scraping-tutorial-python/
- Quick guide for webscraping with `lxml` and `requests` - http://python-guide-pt-br.readthedocs.io/en/latest/scenarios/scrape/
- An open source and collaborative framework for webscraping in a Python environment and quickly writing spiders - https://scrapy.org/