Data Scraping Method: Using Wget in Searches

Otavio Alves de Brito Lucindo da Silva (ot.debrito@gmail.com), Universidade Federal Rural do Rio de Janeiro

Abstract: I present, in summary, a simpler way to collect data from the internet, through a command common to Linux systems, already available in the system terminals.

Keywords: Wget, WebScraping, Digital Humanities.

Data surveys that take place in Brazil are generally more complex to carry out than surveys with the same theme in other countries, such as those belonging to the European Union. One of the main reasons for this is that we do not always have organized data repositories in Brazil in accordance with the FAIR principles (Locatable, Accessible, Interoperable and Reusable), in this way, to obtain all the data necessary for a particular research, it is often necessary to resort to different sources, and thus data scraping and data mining becomes necessary.

I work with Digital Humanities with an emphasis on public policies, most of my work is focused on urban mobility in the metropolitan region of Rio de Janeiro. As my research is focused on cities that are not part of the capital, finding all the necessary data, such as: routes, roads, public transport lines, number of vehicles and other data on social, structural and environmental aspects, is not always so easy. simple, is when I resort to data scraping methods.

Python has many libraries to perform webscraping, however, for a researcher like me trained in the business areas, the use of some tools of this programming language are quite complex. In the search for a simpler way to download data from the web, I came across the Linux command Wget, which has helped me a lot with data extraction, in order to complement the data already existing in repositories.

In one of my works, I needed data about a certain city, which I only found available on a website called Moovit, however, extracting data from the website with Python's Request library, for example, seemed like a very complex job, with the use from Wget, I only needed the following command in the Linux terminal: wget --mirror -p --convert-links -P ./LOCAL http://www.site.com.br so that I had all the data in my computer, requiring only a second Python script to clean the data, using pandas and numpy libraries.

I believe that in terms of data cleaning, mainly, we have many useful tools on the market, such as VantagePoint itself, which “is a powerful text “mining” tool for discovering knowledge in search results in patents and literature databases.”, however as for the collection itself, i.e. on how to extract unorganized data from a website, to organize it in a repository, scraping tools like Request, BeautifulSoup are extremely useful, but generally not as simple as Wget , which, in general, usually asks the user to indicate the desired site.

References

Universidade de Coimbra, Princípios FAIR <https://www.uc.pt/openscience/sobre/acesso-aberto/fair/#:~:text=Os%20Princ%C3%ADpios%20FAIR%20foram%20publicados,do%20ecossistema%20da%20Ci%C3%AAncia%20Aberta.> Acessado em Agosto de 2022.

Wikipédia, Wget <https://pt.wikipedia.org/wiki/Wget> Agosto de 2022.

VantagePoint <https://software.com.br/p/vantagepoint> Acessado em Agosto de 2022.