Web Scraping: A Data Science ability for Statisticians

Alex Sanchez-Pla

http://github.com/ASPteaching/WebScraping_with_R

Web scraping is a general term used to describe data extraction from websites mainly with some degree of automation. While this is one of these tasks that one typically relates with Python and Data science it is not necessarily the case. In this talk I will do a quick overview of a few aspects of web scraping that any statistician can be interested in learning for two main reasons. First because it provides a relatively simple way to access data that we often see to be there (e.g. html tables, or tweets) but think they are too hard to obtain. Second because learning web scraping requires acquiring a variety of abilities, such as XML, regular expressions or APIs that can be very useful in many situations, not only for web scraping. And last, but not least, because it may be really fun and it will get you an applause when you show your Friends how you collected their tweets and made a nice wordcloud of their thoughts.