DLI Accelerated Data Science Teaching Kit

Lecture 2.3 - Popular Scraping Libraries

The Accelerated Data Science Teaching Kit is licensed by NVIDIA, Georgia Institute of Technology, and Prairie View A&M University under the Creative Commons Attribution-NonCommercial 4.0 International License.

Popular Scraping Libraries

Selenium. Supports multiple languages. http://www.seleniumhq.org
Beautiful Soup. Python. https://www.crummy.com/software/BeautifulSoup
Scrapy. Python. https://scrapy.org
JSoup. Java. https://jsoup.org

Important considerations:

Different web content shows up depending on web browsers used
Scraper may need different “web driver” (e.g., in Selenium), or browser “user agent”

Data may show up after certain user interaction (e.g., click a button)
Scraper may need to simulate the actions.
Selenium supports more actions:
http://www.discoversdk.com/blog/web-scraping-with-selenium
Beautiful Soup supports some.

GTx

DEEP
LEARNING
INSTITUTE

PRAIRIE VIEW
A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

Thank You