Web Scraping

PRESENTED BY

DAVID SELASSIE OPOKU

@sdopoku & @schoolofdata

29 August 2015

SCHOOL OF DATA

mest
meltwater entrepreneurial
school of technology

About School of Data

SCHOOL OF DATA

Who Are You?

1. Name
2. Where you are from
3. Background and interest
4. One random fact about you

SCHOOL OF DATA

Outline

I'm Doing All the Talking

1. Why [Open] Data & The Data Pipeline
2. What is and Why Data Scraping?
3. Best Practices & Tools

DIY Time

4. 3 Cases of Scraping
5. Reference Resources

SCHOOL OF DATA

Why [Open] Data ?

SCHOOL OF DATA

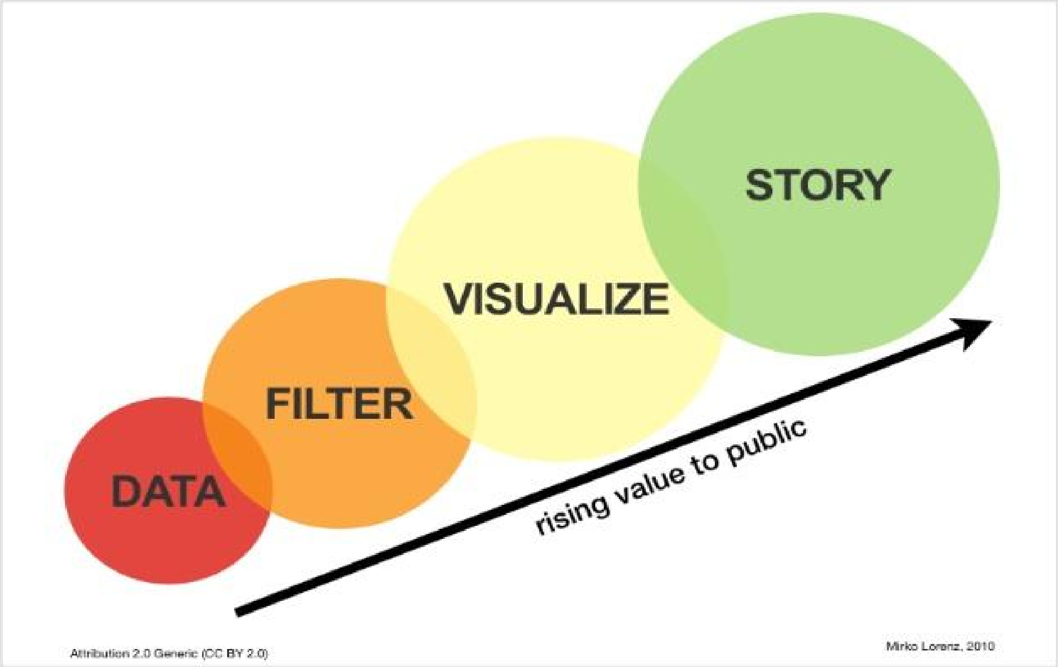
Group Activity: 15 mins

1. Why Data?
2. Why Open Data?
3. Write down some data buzz words you have heard recently

SCHOOL OF DATA

Data Pipeline

SCHOOL OF DATA



Attribution 2.0 Generic (CC BY 2.0)

Mirko Lorenz, 2010

Target Audience

SCHOOL OF DATA

This should be useful to ...

- Non-tech-savvy data enthusiasts
- Advanced data enthusiasts
- Web developers & data publishers
- Data journalists

SCHOOL OF DATA

What is & Why Data Scraping ?

SCHOOL OF DATA

Data Scraping: what is it ?

scrape [*verb* \ˈskrāp\ ]

: to remove from a surface by usually repeated strokes of an edged instrument

: to collect by or as if by scraping —often used with *up* or *together* <*scrape* up the price of a ticket>

- **Merriam Webster**

*“The **transformation of unstructured data** on the web, typically in HTML format, **into structured data** that can be **stored and analyzed** in a central local database or spreadsheet.”*

- **Wikipedia (web scraping)**

SCHOOL OF DATA

When should you scrape data ?

CSV JSON

Machi le data

- PDF Data
- HTML data

Best Practices

SCHOOL OF DATA

Best Practices For Scrapers

1. Scraping is not scary!
    a. Use existing tools
2. Use a modern and friendly browser
    a. Chrome, Firefox, Opera, Safari
    ~~b. Avoid Internet Explorer~~
3. Map out the process
    a. Where does scraping fit in?

Best Practices For Data Publishers

1. Have a consistent structure
    a. Websites
    b. PDFs
2. Always think about your data end users
    a. Before, during & after publishing

SCHOOL OF DATA

Steps

1. Map out the process/pipeline for your data project

2. Identify your data source (website, PDF, API?)

3. Decide on storage format for your scraped data
   a. CSV file, Spreadsheet, Google docs
   b. Database

4. Select scraping tool

5. Verify and Clean data

Tools

SCHOOL OF DATA

Tools: Web Browsers

Tools: Scraping Apps

1. Point and click
   a. Scraper Google Chrome extension
   b. Webscraper.io, Import.io, Kimono Labs,
   c. Tabula (PDF)
2. Programming (Python & Rlibraries)
   a. Pattern (PDF and HTML)
   b. Beautiful Soup
   c. rvest
   d. Scrapy

SCHOOL OF DATA

Tools: Storage & Sharing

1. Google Spreadsheets
2. Github
3. Datahub.io

SCHOOL OF DATA

Time To Try It Out

SCHOOL OF DATA

Tabula PDF Scraper

Sometimes, the data you need can only be found in a PDF. This is where the Tabula PDF scraper tool can be useful.

The Tabula website provides great instructions on what Tabula is, and how to install and use it for Windows, Mac and Linux operating systems.

NB: Tabula on works for text data in tables and also not for scanned PDFs.

SCHOOL OF DATA

Tabula

Tabula helps you liberate data tables trapped inside PDFs

HTML Scraping

SCHOOL OF DATA

Scraper Chrome Extension

NB: This tool only works for Google Chrome browser

Installation

1. Make sure you have installed Google Chrome
2. Open up Chrome and visit the Web Store at https://chrome.google.com/webstore
3. Search for "***scraper extension***" in the search bar on the top left corner of page
4. The scraper tool is the 1st one under the **Extensions** section
5. Click on "**Add to Chrome**" to download & install into Chrome

Scraper Chrome Extension

Usage

1. Open up an HTML page with a table of data you want to scrape

   Eg: List of Africa sovereign states from Wikipedia
2. Find the HTML data in the article
3. Starting from inside the 1st row, highlight a couple of rows
4. Right click & select the **“Scrape similar”** option
5. This will open up a window with the data from the table
6. Copy the data to the clipboard or save into Google Spreadsheet.

SCHOOL OF DATA

Webscraper.io

This is another “Point and click” web scraping tool but with some advanced capabilities to scrape from paginated and nested websites.

The webscraper.io page give a wealth of information about the tool and also have great video tutorials which you should check out at http://webscraper.io/tutorials

Resources - Readings and Tools

1. Five data scraping tools for would-be data journalists
2. Making data on the web useful: scraping
3. Liberating HTML Data Tables
4. BeautifulSoup Python Library
5. Pattern Python Library
6. Scrapy Python Library
7. Datahub
8. Import.io & Kimono
9. Webscraper.io
10. Tabula
11. rvest R package

SCHOOL OF DATA