Scraping the Web

Akitaka Matsuo

Essex IADS

Scraping the web: what?

An increasing amount of data is available on the web:

- Speeches, sentences, biographical information…
- Social media data, newspaper articles, press releases…
- Geographic information, conflict data…

These datasets are often provided in an **unstructured format**.

Web scraping is the process of extracting this information automatically and transforming it into a **structured dataset.**

Scraping the web: why?

Copy & pasting is time-consuming, boring, prone to errors, and impractical for large datasets

In contrast, automated web scraping:

1. Scales well for large datasets
2. Is reproducible
3. Involved adaptable techniques
4. Facilitates detecting and fixing errors

When to scrape?

1. Trade-off between your time today and your time in the future. **Invest in your future self!**
2. Computer time is cheap; human time is expensive

Scraping the web: two approaches

Two different approaches:

1. **Screen scraping**: extract data from source code of website, with html parser and/or regular expressions
   - `urlopen` + `BeutifulSoup` in Python
2. **Web APIs** (application programming interfaces): a set of structured http requests that return JSON or XML data
   - `urlopen` to construct API requests, then `xml` or `json` package to parse
   - Packages specific to each API: wwo-hist, World-Bank-Data, Tweepy, uk-covid19