ECPR Methods Summer School: Automated Collection of Web and Social Data

Pablo Barberá

London School of Economics
pablobarbera.com

Course website:
pablobarbera.com/ECPR-SC104

Scraping the web


CANT SCRAPE WEBSITE
WEBSITE USES AJAX

Advanced scraping

Selenium:

- General idea: browser control to scrape dynamically rendered web pages
- Originally developed for web testing purposes
- R will launch a browser session and all communication will be routed through that browser session.
- `phantomJS`: headless browser (will not display website)
- Capabilities: complete forms, write text, click on buttons or area of website, navigate to new URL...

Scraping newspaper websites

RSS feeds

- Really Simple Syndication, originally developed as a way to regularly check for new content on sites
- Includes list of entries (with some more information) and when they were updated
- Written in XML format (eXtensible Markup Language)
- Example: The Guardian RSS feed

Social event

Save the date:
Wednesday Aug. 1st, 6pm
Location TBA