Web scraping with R

Will Lowe
wlowe@princeton.edu
Q-APS, Department of Politics

10:00-12:00 and 1:00-3:00, May 23, 2018

This short workshop focuses on using R to extract content from the web. This includes getting content from static web pages, or by filling in web forms and harvesting their results. We will also describe how to use common styles of web APIs, as provided e.g. by Twitter or ProPublica, from R.

Prerequisites

We will assume a moderate level of R competence. We will asume that you have a good grasp of R's data and control flow structures, and that you are comfortable writing functions. If you have taken the Q-APS R Camp this is quite sufficient. We assume no prior experience with web data.

Setup

You should have R and RStudio on your machine and the ability to install R packages. Course materials will be available on Blackboard.

Mac users may find it helpful to install the 'Developer Tools'. To do so, open a Terminal window, type `make` and follow any prompts to download 'commandline tools'. If you get message that there are 'no targets specified', then these tools are probably already installed.

R packages can be installed as needed during the workshop. We will mostly work with `rvest`.

Structure

The course has three parts, each with a practical exercise

Day	Session	Details
Wednesday	10-12pm	How webpages are structured How servers deliver them Identifying and extracting page content with CSS tags and XPath
	1-3pm	Introduction to web forms Driving web forms from R Harvesting etiquette
	3pm+	Working with web-based APIs, e.g. Twitter, ProPublica Authentication, REST interfaces, and JSON data Common scraping problems and solutions