Webscraping in R using Rvest

Loren Collingwood

To harvest text, the webscraping skillset is a must. This vignette introduces webscraping in R using the package Rvest. When I was growing up, we used Python to webscrape but now that is less necessary. In my experience, you cannot always webscrape everything, and a one-size fits all

Step 1

Download the selector gadget chrome extension.

Step 2

Install then load the Rvest package.

```
    # Install R package if you have not done so #
    #install.packages("rvest")

    # Load package into current R session #
    library("rvest")
#> Loading required package: xml2
```

Step 3

Select a website you want to scrape and investigate.

1. Investigate whether the particular information you want from the website can be scraped.
2. Is the information presented across a series of pages?
3. Is the information possibly in a table?
4. Are there multiple bits of information you want to scrape? i.e., title, author, date, main body of text.
5. Think about how you want the data structure to look like when webscraping is complete. Do you want the text in single .txt files, a single .csv file with columns for metadata?

Step 4

Open the website (if not already open). Click on the selector gadget and begin looking for unique html tags that identify what you want to scrape

Example

Core Civic or Corrections Corporation of America (CCA) is a private corrections firm publicly traded on the NYSE. The company produces newsletters to shareholders publicly available on its website. Let's scrape these.

Summary

This vignette has provided a brief overview of the workflow for using webscraping in R.

- `resolve_missing_vals()`