MAT381E-Week 7: Introduction to Web Scraping

Gül İnan

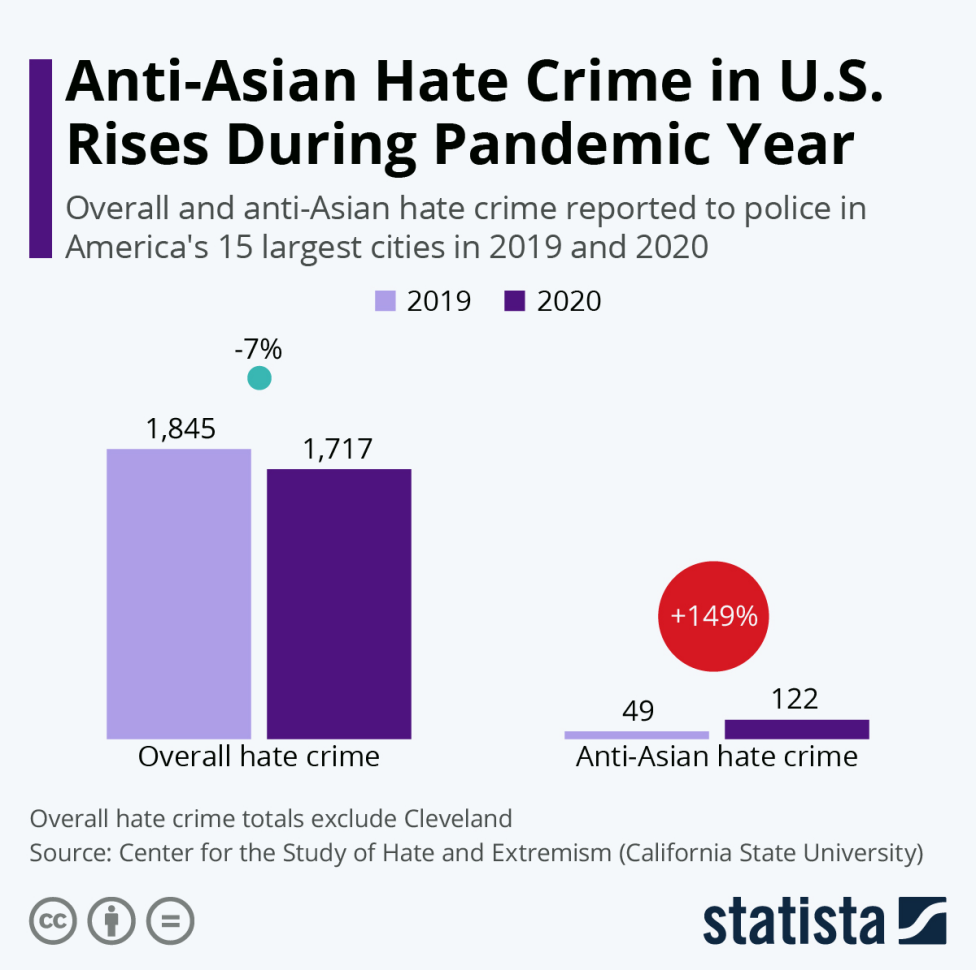
Department of Mathematics
Istanbul Technical University

April 11, 2021

Outline

- Motivation.
- What is `Web Scraping`?
- `HTML` basics.
- Web scraping with `rvest` package.
- Ethical issues.
- 01-web_scraping.Rmd (next week).

Motivation



Source

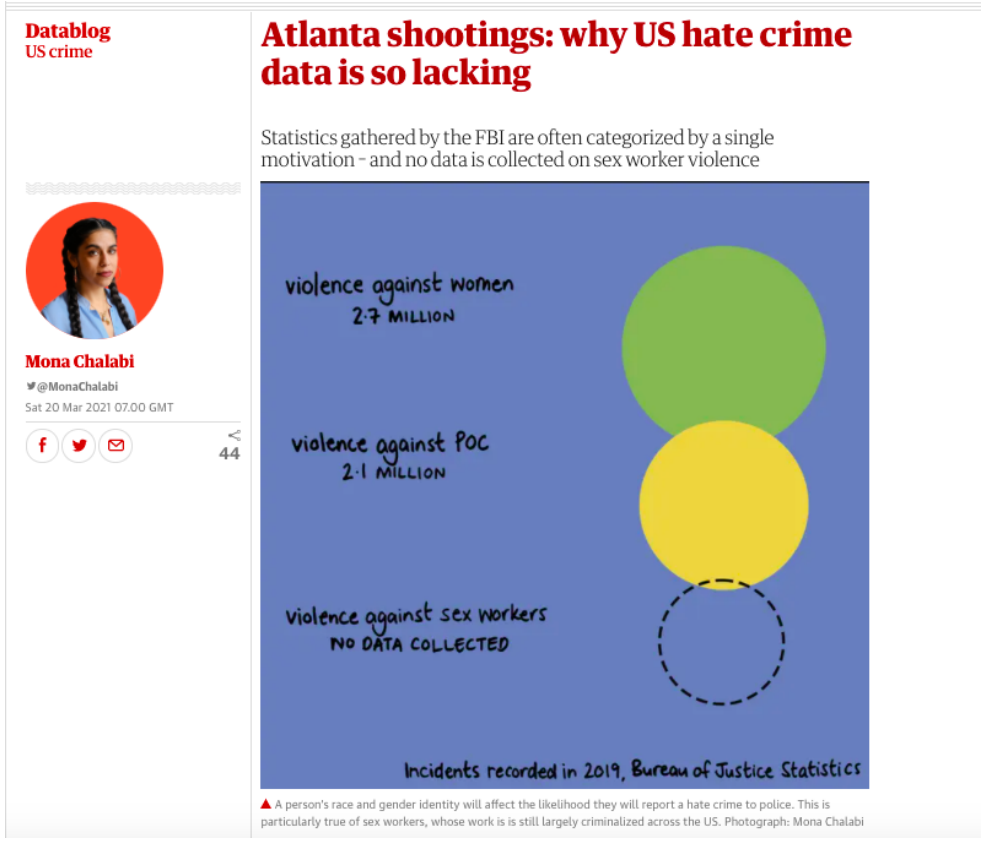
- "A survey of police reports by the Center for the Study of Hate and Extremism at California State University confirmed that racially motivated crimes against those of Asian descent in the U.S. have risen in the pandemic year of 2020. **While hate crimes against Asians still make up a smaller fraction of all hate crimes reported in America's 15 largest cities, their number rose from 49 in 2019 to 122 in 2020.**"
- "Stop AAPI Hate said yesterday that verbal harassment was the most common incident recorded by them at 68 percent of all cases, followed by deliberate shunning (20 percent of cases) and physical attacks (11 percent of cases)."

What is a hate crime?

- "A hate crime, by definition, includes a motivation rooted in bias, according to the Justice Department. Bias can be based on a victim's perceived or actual race, color, religion, national origin, sexual orientation, gender, gender identity or disability. "
- "The Justice Department adds: "Hate crimes have a devastating effect beyond the harm inflicted on any one victim. They reverberate through families, communities, and the entire nation, as **others fear that they too could be threatened, attacked, or forced from their homes, because of what they look like, who they are, where they worship, whom they love, or whether they have a disability.**"

Source

Lacking Hate Crime Data



Datablog
US crime

Atlanta shootings: why US hate crime data is so lacking

Statistics gathered by the FBI are often categorized by a single motivation - and no data is collected on sex worker violence

Mona Chalabi
@MonaChalabi
Sat 20 Mar 2021 07.00 GMT

44

▲ A person's race and gender identity will affect the likelihood they will report a hate crime to police. This is particularly true of sex workers, whose work is is still largely criminalized across the US. Photograph: Mona Chalabi

Source

- "This, of course, ignores the possibility that someone might be motivated by racial hatred and sexism."
- "Unfortunately, most statistics make the same assumption. Hate crime data that is gathered by the FBI is often categorized according to **a single motivation** (such as religion, sexual orientation, race/ethnicity, gender identity). Less than 3% of the hate crimes that were reported in 2019 recorded **multiple biases.**"
- "**Reality is obviously much more complex than these numbers capture.** Things get even more complicated when you consider reporting rates. A person's race and gender identity will affect the likelihood that they will report a hate crime to the police."

Motivating Data

- The data we need to answer a question may not always come in a spreadsheet and be ready for us to read. Sometimes, data can be available on the web.
- For example, following Wikipedia page illustrates **Hate crime statistics by bias motivation in the US** in a `html` table:

Victims per Year by Bias Motivation[120]

Department of Justice / FBI Hate Crimes Statistics

Bias Motive	1995	1996[123]	1997[124]	1998[125]	1999[126]	2000[127]	2001[128]	2002[129]	2003[130]	2004[131]	2005[132]	2006[133]	2007[134]	2008[135]	2009[136]	2010[137]	2011[138]	2012[13
Race	6,438	6,994	6,084	5,514	5,485	5,397	5,545	4,580	4,754	5,119	4,895	5,020	4,956	4,934	4,057	3,949	3,645	3,467
Race/Ethnicity/Ancestry																		
Religion	1,617	1,535	1,586	1,720	1,686	1,699	2,118	1,659	1,489	1,586	1,405	1,750	1,628	1,732	1,575	1,552	1,480	1,340
Sexual Orientation	1,347	1,281	1,401	1,488	1,558	1,558	1,664	1,513	1,479	1,482	1,213	1,472	1,512	1,706	1,482	1,528	1,572	1,376
Ethnicity/National Origin	1,044	1,207	1,132	956	1,040	1,216	2,634	1,409	1,326	1,254	1,228	1,305	1,347	1,226	1,109	1,122	939	866
Disability	*unknown*	*unknown*	12	27	23	36	37	50	43	73	54	95	84	85	99	48	61	102
Gender	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknov*
Gender Identity	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknown*	*unknov*
Single-Bias	**10,446**	**11,017**	**10,215**	**9705**	**9,792**	**9,906**	**11,998**	**9,211**	**9,091**	**9,514**	**8,795**	**9,642**	**9,527**	**9,683**	**8,322**	**8,199**	**7,697**	**7,151**
Multiple-Bias	23	22	40	17	10	18	22	11	9	14	9	10	8	8	14	9	16	13
Total	**10,469**	**11,039**	**10,255**	**9,722**	**9,802**	**9,924**	**12,020**	**9,222**	**9,100**	**9,528**	**8,804**	**9,652**	**9,535**	**9,691**	**8,336**	**8,208**	**7,713**	**7,164**

Notes: The term *victim* may refer to a person, business, institution, or society as a whole. Though the FBI has collected UCR data since 1992, reports from 1992-1994 are not available on the FBI website. Single-bias victim totals have been calculated for 1995-1998. *Race* and *Ethnicity/National origin* were merged together starting in 2015.

Web Scraping

- **Web scraping** or **web harvesting** are the terms used to describe the process of extracting data from a website.
- The reason we can do this is because the information used by a **browser** to render **web pages** is received as **text** from a server.
- This text is computer code written in **hyper text markup language** (HTML).
- To see the `HTML` source code for a web page we can visit the page on the *browser*, then we can use the *View Page Source* tool to see it.
- Because `HTML` code is accessible, we can download the `HTML` files, import it into `R`, and then write code to extract the information we need from the page.

- To get an idea of how `HTML` code works, here we show a few lines of code from the Wikipedia page that provides information on US hate statistics:

```
</p>
<h2><span class="mw-headline" id="Prevalence_of_hate_crimes">Prevalence of hate crimes</span><span class="mw-editsection"><span class="mw-editsection-bracket">[</span><a href="/w/index.php?title=Hate_crime_laws_in_the_United_States&amp;action=edit&amp;section=14" title="Edit section: Prevalence of hate crimes">edit</a><span class="mw-editsection-bracket">]</span></span></h2>
<p>The DOJ and the FBI have gathered statistics on hate crimes reported to law enforcement since 1992 in accordance with the <a href="/wiki/Hate_Crime_Statistics_Act" title="Hate Crime Statistics Act">Hate Crime Statistics Act</a>. The FBI's <a href="/wiki/Criminal_Justice_Information_Services_Division" class="mw-redirect" title="Criminal Justice Information Services Division">Criminal Justice Information Services Division</a> has annually published these statistics as part of its <a href="/wiki/Uniform_Crime_Reports" title="Uniform Crime Reports">Uniform Crime Reporting</a> program. According to these reports, of the over 113,000 hate crimes since 1991, 55% were motivated by racial bias, 17% by religious bias, 14% sexual orientation bias, 14% ethnicity bias, and 1% disability bias.<sup id="cite_ref-122" class="reference"><a href="#cite_note-122">&#91;122&#93;</a></sup> <a href="/wiki/David_Ray_Hate_Crimes_Prevention_Act" title="David Ray Hate Crimes Prevention Act">David Ray Hate Crimes Prevention Act</a>
</p><p>Please note that the figures in the table below do not contain data from all reporting agencies every year. 2004 figures covered a population of 254,193,439, 2014 covered 297,926,030.
</p>
```

Prevalence of hate crimes [edit]

The DOJ and the FBI have gathered statistics on hate crimes reported to law enforcement since 1992 in accordance with the Hate Crime Statistics Act. The FBI's Criminal Justice Information Services Division has annually published these statistics as part of its Uniform Crime Reporting program. According to these reports, of the over 113,000 hate crimes since 1991, 55% were motivated by racial bias, 17% by religious bias, 14% sexual orientation bias, 14% ethnicity bias, and 1% disability bias.[122] David Ray Hate Crimes Prevention Act

Please note that the figures in the table below do not contain data from all reporting agencies every year. 2004 figures covered a population of 254,193,439, 2014 covered 297,926,030.

- Once we look at the full `HTML` source code, we can actually see the text and data along with `HTML` codes.
- We can also see **a pattern** of how it is stored. If you know `HTML`, you can write programs that leverage knowledge of these patterns to extract what we want.
- We also take advantage of a language widely used to make web pages look "pretty" called Cascading Style Sheets (CSS).

HTML basics

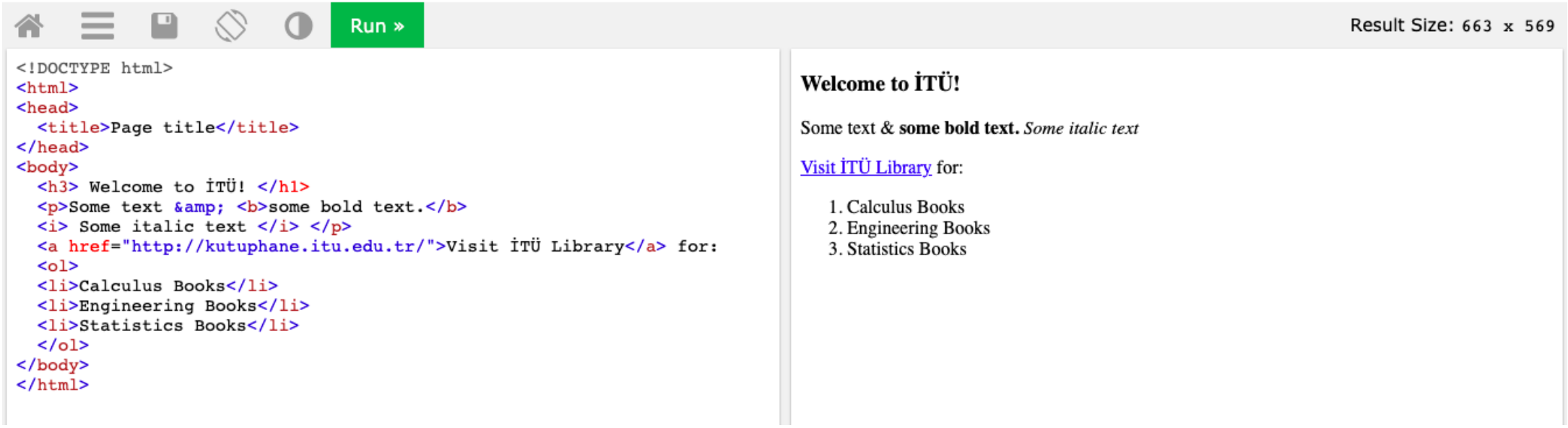
- All `HTML` documents must start with a document type declaration: `<!DOCTYPE html>`.
- Every `HTML` page itself must be in an `<html>` element, and it must have two children: `<head>`, which contains document metadata like the page title, and `<body>`, which contains the content you see in the browser.

```
<!DOCTYPE html>
<html>
<head>
  <title>Page title</title>
</head>
<body>
  <h1> Welcome to İTÜ! </h1>
  <p>Some text &amp;amp; <b>some bold text.</b>
  <i> Some italic text </i> </p>
  <a href="http://kutuphane.itu.edu.tr/">Visit İTÜ
  <ol>
  <li>Calculus Books</li>
  <li>Engineering Books</li>
  <li>Statistics Books</li>
  </ol>
</body>
</html>
```

- Each `HTML` element has a hierarchical structure which consist of a start tag (e.g. `<tag>`), optional attributes (`id='first'`), an end tag (like `</tag>`), and contents (everything in between the start and end tag).
- Block tags like `<h1>` (most important heading 1), `<p>` (paragraph), and `<ol>` (ordered list), `<li>` (list item) form the overall structure of the page.
- Inline tags like `<b>` (bold), `<i>` (italics), and `<a>` (links) formats text inside block tags.
- On the left: The `<a>` tag defines a hyperlink. The `href` **attribute specifies the URL of the page the link goes to**.

- Note: Since `<` and `>` are used for start and end tags, we cannot use them directly.
- Instead we have to use the `HTML` escapes `>` (greater than) and `<` (less than).
- And of couse, since those escapes use `&`, if we want a literal ampersand (and) we have to escape it as `&amp`.

- Let's try out our `HTML` code at WWW3 school:



- More on HTML.

- Some elements, like `<img>` cannot have children. These elements depend solely on attributes for their behavior.

```
<img src='logo/rvest.jpg' width="400" height="400">
```

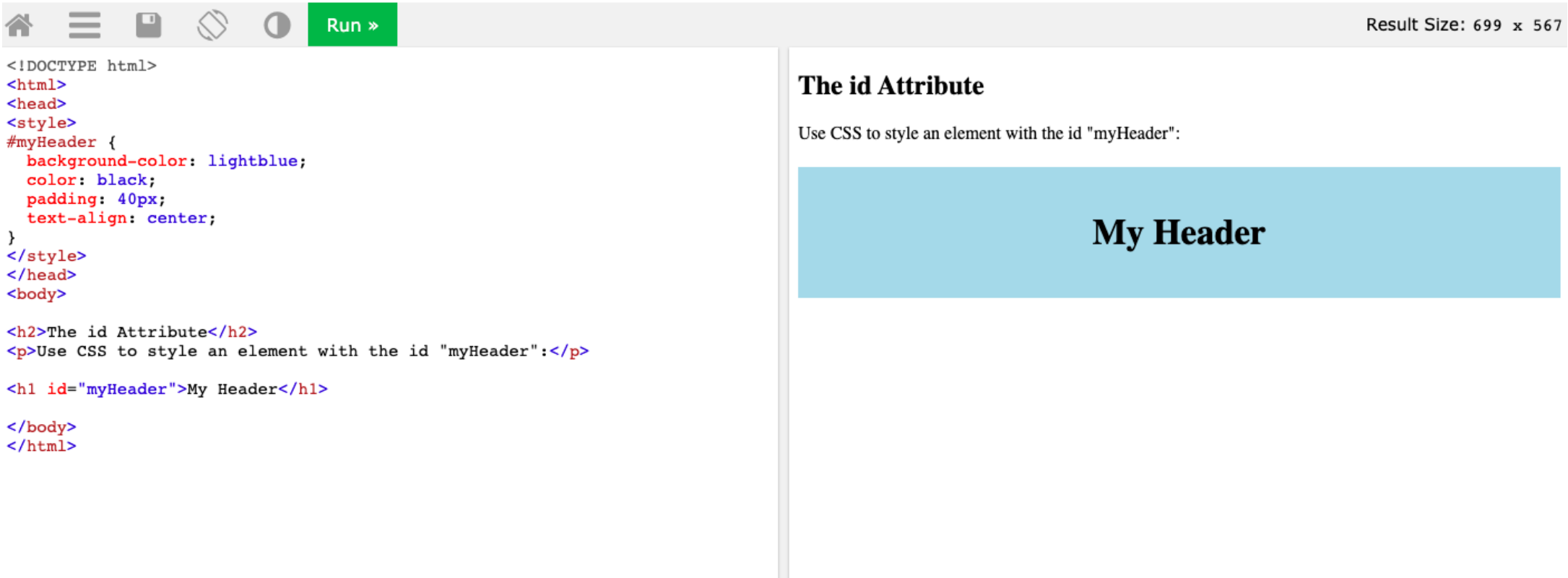
- Here, `src` attribute specifies the path (URL) to the image; `width` and `height` define the `width` and `height` of the image in pixels.

Attributes

- Sometimes, the start tags of `HTML` elements can have **named attributes** which look like Content .`
- Two of the most important named attributes are `id` and `class`, which are used in conjunction with `CSS` to **control the visual appearance** of the page. These are often useful when scraping data off a page.
- Note that attributes are always specified in the start tag.

Using id attribute

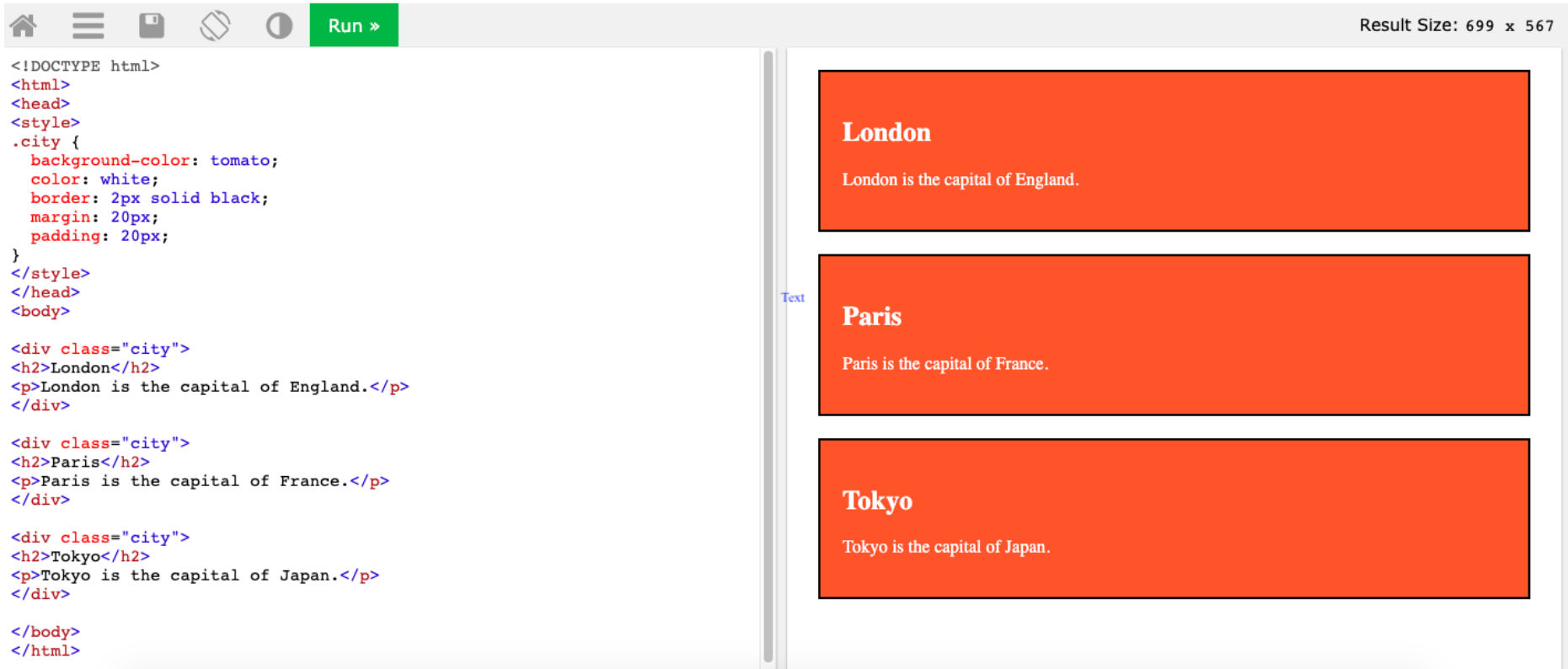
- The `id` attribute is used to point to a specific style declaration in a style sheet and the value of the `id` attribute must be **unique** within the `HTML` document.
- The syntax for `id` is: write a hash character (`#`), followed by an `id name`. Then, define the CSS properties within curly braces `{}`.



Source1 and Source2

Using class attribute

- The `class` attribute is often used to point to a class name in a style sheet. Multiple `HTML` elements can share the same class.
- The syntax for `class` is: write a period character (`.`), followed by an `class name`. Then, define the CSS properties within curly braces `{}`.



Source1 and Source2

- Note that main difference between `id` and `class` attribute is that `id` is unique in a page and can only apply to **at most one HTML element**, while `class` attribute can be applied to **multiple HTML elements**.

Rvest

The rvest package

- The `rvest` package provides web harvesting tools within `tidyverse` ecosystem.

```
# rvest is not within the core tidyverse ecosystem
# library(tidyverse) will not load rvest package
# load rvest package by library(rvest) call specifically
library(rvest)
```

- The `rvest` manual tells us that it depends on a few other packages including `xml2`. This enables us to use functions available in these packages as well.

- Here are basic `rvest` functions:

Function	Description
`read_html()`	takes a string that can be either a path, a url and then creates a HTML document from a webpage.
`html_nodes()`	select specified nodes with the specified tags from the HTML document.
`html_table()`	extract table, to be used after `html_nodes()`.
`html_text()`	extract text within tags, to be used after `html_nodes()`.
`html_attr()`	extract the value of attribute, to be used after `html_nodes()`.

- The first step in using this package is to import the web page, you are interested in, into R.

```
# Use `read_html()`: to read HTML data from a url or character string into R.
url <- "https://en.wikipedia.org/wiki/Hate_crime_laws_in_the_United_States"
h   <- read_html(url)
h
```

```
#> {html_document}
#> <html class="client-nojs" lang="en" dir="ltr">
#> [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
#> [2] <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject  ...
```

- Note that the entire Wikipedia webpage is now contained in `h`:

```
h
```

```
#> {html_document}
#> <html class="client-nojs" lang="en" dir="ltr">
#> [1] <head>\n<meta http-equiv="Content-Type" content="text/html; charset=UTF-8 ...
#> [2] <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject  ...
```

- The `h` object is a *list* (`R` data type) and the items in the `h` object correspond to the basic document structure of an `HTML` document.
- Displaying the `h` object shows that the first item in the *list* is `head` and the second item is `body`.
- Note that these items include the basic component of the `HTML` document, in other words, the *text, links*, and HTML "stuff" were scraped from the web page.
- Specifically this stuff is found in the *body* element of the `h` *list*.

Extract a table

- Now, question is "how do we extract the table from the object `h`?"
- Remember that `HTML` code has a hierarchical tree structure. The different parts of an `HTML` code, often defined with a message in between `<` and `>` are referred to as *nodes* (in other words, tags).
- When we know that the information is stored in an `HTML table`, we can see this in the `HTML code` with `<table>` tags.
- To extract a table from the `h` *list*, then we need to gather all the `HTML` code within the `<table>` tags in the `h` *list*.
- You can learn more about the `<table>` tag structure from HTML documentation.

- The `rvest` package includes functions to extract nodes of an `HTML` document: `html_nodes()` extracts all nodes of different type and `html_node` extracts the first one. To extract all tables we use:

```
tab <- h %>%
       html_nodes("table")
# note that in HTML source code there are currently 5 tables!..
# pages are up to change!..
tab
```

```
#> {xml_nodeset (5)}
#> [1] <table class="box-Cite_check plainlinks metadata ambox ambox-content" rol ...
#> [2] <table class="wikitable">\n<caption>\n</caption>\n<tbody>\n<tr>\n<th>Stat ...
#> [3] <table class="wikitable" style="margin: 1em auto 1em auto">\n<caption>\n< ...
#> [4] <table class="wikitable" style="margin: 1em auto 1em auto">\n<caption>\n< ...
#> [5] <table class="box-Unbalanced plainlinks metadata ambox ambox-content ambo ...
```

- Now, instead of the entire web page, we just have the `HTML` code for the **tables only**:

- But we want the table titled "Victims per Year by Bias Motivation" on the page. - Looking at the output above it looks like the table index is [3]. To extract just the third table - the table with the data we are interested in - we can type the following:

```
tab <- h %>%
        html_nodes("table") %>% .[3]
# subsetting with square brackets while piping: .[]
tab
```

```
#> {xml_nodeset (1)}
#> [1] <table class="wikitable" style="margin: 1em auto 1em auto">\n<caption>\n< ...
```

- We are not quite there yet because this is clearly not a tidy dataset, not even a data frame.
- In the code above you can definitely see a pattern and writing code to extract just the data is very doable. In fact, `rvest` includes a function just for converting `HTML` tables into data frames:

```
table_victim <- tab %>%
                  html_table %>% .[[1]]  #returns a list
#View(table_victim)
```

- We are now much closer to having a usable data table. Change the column names properly, replace "unknown" and empty spaces with NA, then remove the commas and turn characters into numbers.

```
library(dplyr)
table_tidy <- table_victim  %>%
                setNames(c("Bias Motive", paste(c(1995:2018), sep=""))) %>%
                na_if("unknown") %>%
                na_if("") %>%
                ###mutate_at(vars("1995":"2018"), as.numeric)  #did not work!help needed
# https://stackoverflow.com/questions/46787515/remove-commas-from-character-vectors-based-on-specif
                mutate_at(vars("1995":"2018"), funs(as.numeric(gsub(',', '', .))))
#View(table_tidy)
```

```
library(kableExtra)
kbl(cbind(table_tidy, table_tidy)) %>%
  kable_paper() %>%
  scroll_box(width = "100%", height = "400px")
```

Bias Motive	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	20
Race	6438	6994	6084	5514	5485	5397	5545	4580	4754	5119	4895	5020	4956	4934	4057	3949	3645	34
Race/Ethnicity/Ancestry	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
Religion	1617	1535	1586	1720	1686	1699	2118	1659	1489	1586	1405	1750	1628	1732	1575	1552	1480	13
Sexual Orientation	1347	1281	1401	1488	1558	1558	1664	1513	1479	1482	1213	1472	1512	1706	1482	1528	1572	13
Ethnicity/National Origin	1044	1207	1132	956	1040	1216	2634	1409	1326	1254	1228	1305	1347	1226	1109	1122	939	8
Disability	NA	NA	12	27	23	36	37	50	43	73	54	95	84	85	99	48	61	1
Gender	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N
Gender Identity	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	N

Exract Text

Data Code

- Let's assume that you want to extract the following unordered list at the Department of Justice:

```
knitr::include_graphics('images/offense.png')
```

Offenses by Crime Category

Among the 8,552 hate crime offenses reported:

- Crimes against persons: 64.4%
- Crimes against property: 32.8%
- Crimes against society: 2.8%

Exract image URL

- Let's say we would like to import the image of "ortanca" at https://www.bitkivt.itu.edu.tr/foto/Hydrangea_macrophylla_c%c4%b1cek.sem%c4%b1ha.jpg to the R.

```
image  <- read_html("https://www.bitkivt.itu.edu.tr/vt/report.php?sor=665")
image_url <- image %>%
              html_nodes("img") %>%
              html_attr("src") %>% .[3] #we need third url
```

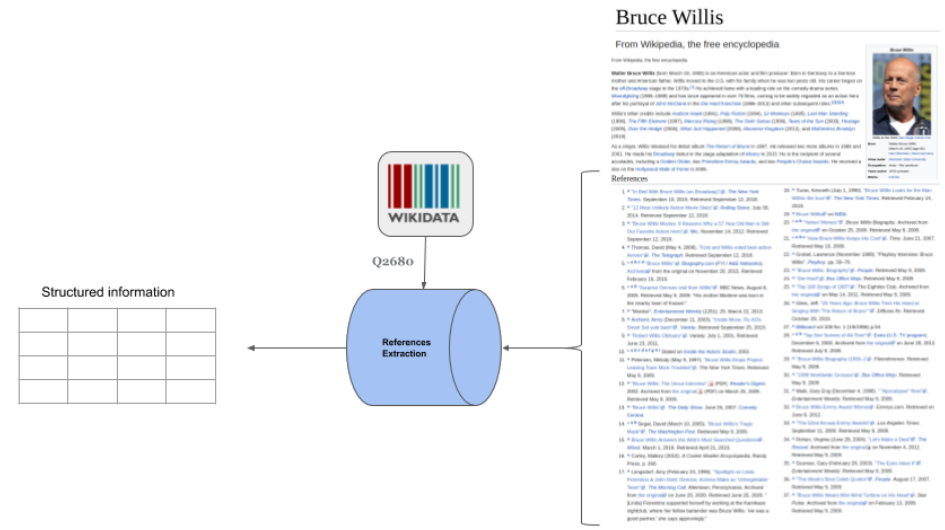
```
#library magick is for image editing (reading, writing, and joining).
library(magick)
magick::image_read(image_url)
```

News

misinfocon.com/scribes-reference-api-enables-users-to-access-wikipedia...

Scribe's Reference API enables users to access Wikipedia references

Lucie-Aimée Kaffee Follow
Mar 25 · 4 min read

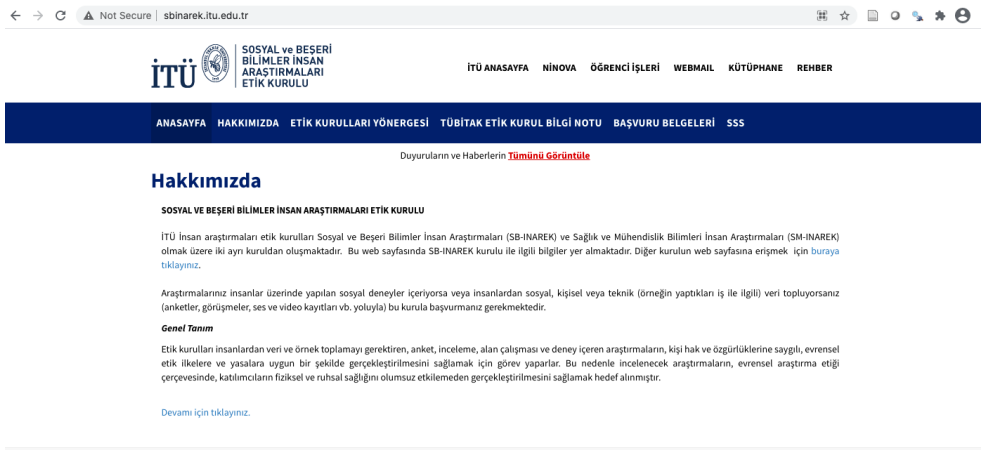


- Scribe says that:
  - "We, therefore, started the Scribe credibility API. The goal was to make the Wikipedia references not only accessible to anyone but also queryable. We implemented this in two steps: (1) extracting Wikipedia references, and (2) setting up an API to query the references."
  - "We extract Wikipedia references from the Wikipedia dump and enrich it with Wikidata information, such as the entity ID in Wikidata. This data is saved as structured data in the database. We focus on online references, i.e., references that include a URL."

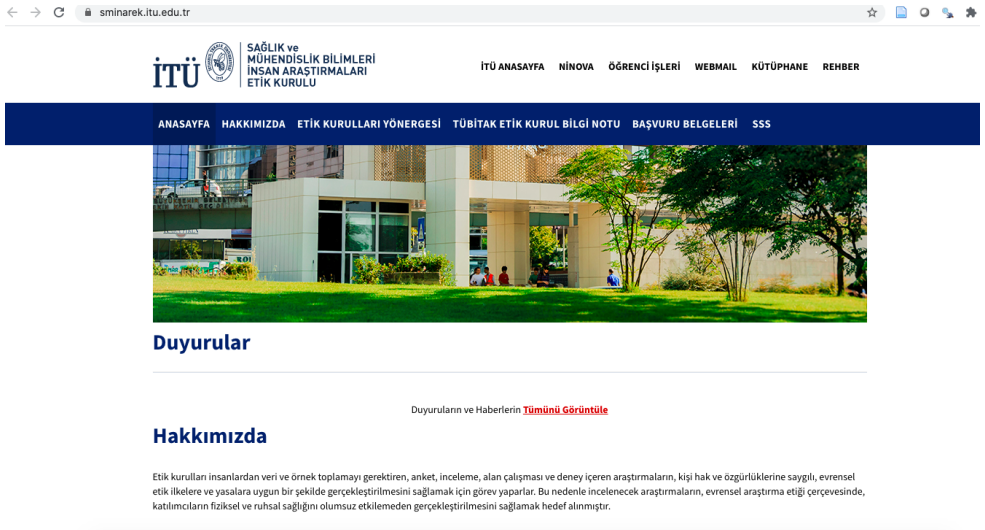
Ethical considerations

- Legal Concerns:
  - If internet data is publicly available (e.g., tweets from a public Twitter account), it is **generally considered legal** to collect this data.
  - Research that involves human participants (e.g., surveys, interviews, blood draws) needs to be approved by the Institutional Ethics Committee.

- "İTÜ İnsan araştırmaları etik kurulları Sosyal ve Beşeri Bilimler İnsan Araştırmaları (SB-INAREK) ve Sağlık ve Mühendislik Bilimleri İnsan Araştırmaları (SM-INAREK) olmak üzere iki ayrı kuruldan oluşmaktadır."



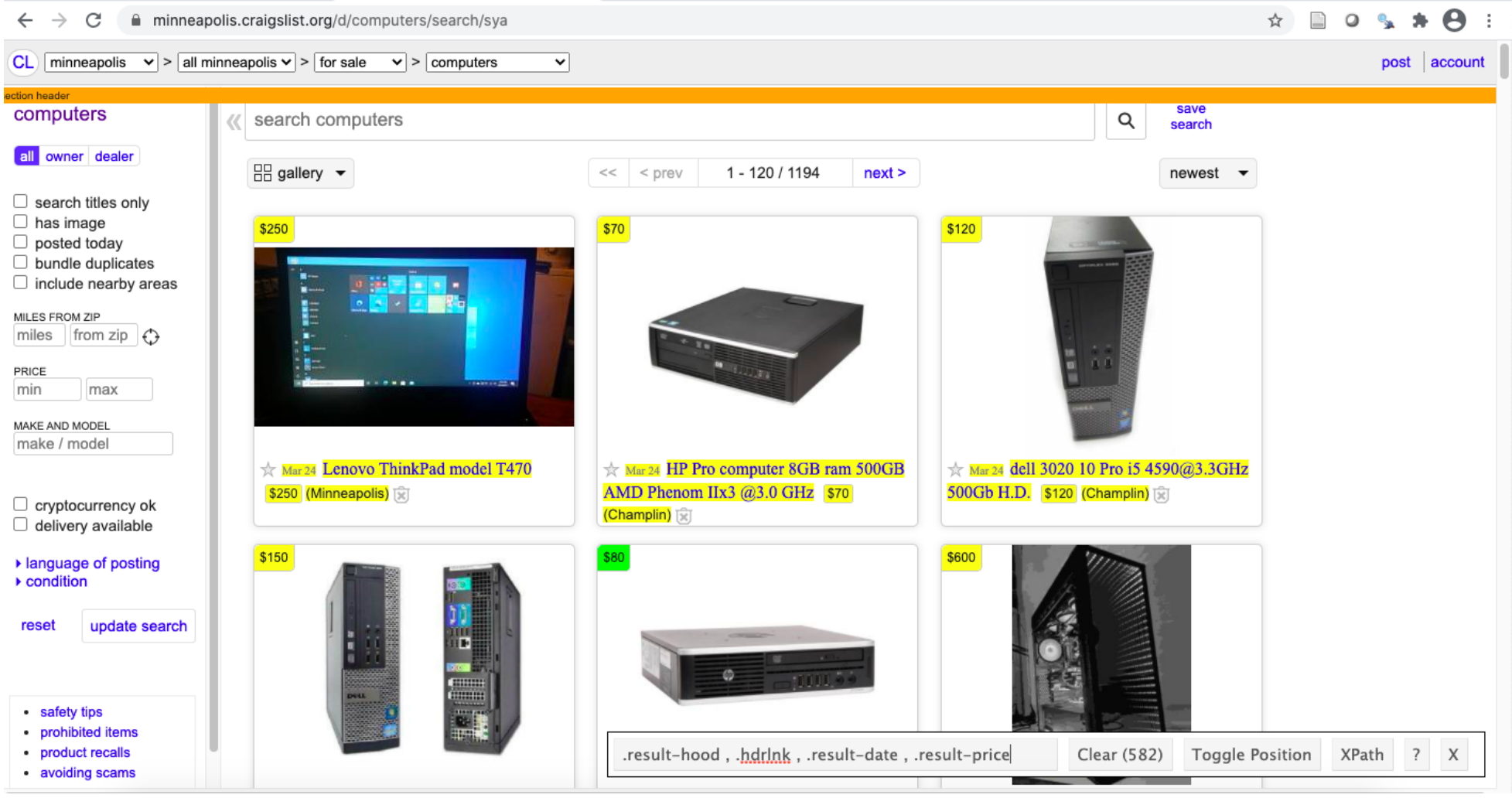
Source



Source

- But it is still not certain whether research about publicly available internet data require Institutional Ethics Committee approval or not.
- User Ethics:
  - According to this information:
    "Just because something is legal does not mean it is ethical. Collecting, sharing, and publishing internet data created by or about individuals can lead to unwanted public scrutiny, harm, and other negative consequences for those individuals. There is no single, simple answer to the many difficult questions raised by internet data collection. It is important to develop an ethical framework that responds to the specifics of your particular research project or use case (e.g., the platform, the people involved, the context, the potential consequences, etc.)."

- **Hands-on example:** Visit `01-web_scraping.Rmd` file for data harvesting from craiglist.



- More on web scraping:
  - https://www.r-bloggers.com/2020/01/web-scraping-with-rvest-astro-throwback/
  - https://www.storybench.org/scraping-html-tables-and-downloading-files-with-r/
  - https://cran.r-project.org/web/packages/RSelenium/vignettes/basics.html.

Attributions

- rvest.
- Data Science Labs.
- Ethics.
- CSS Selectors.