Lecture 10 Web Scraping

Pierre Biscaye Université Clermont Auvergne

Data Science for Economics

Note: Materials for this lecture are drawn from the UC Berkeley D-Lab's Python Web Scraping course.

Why web scraping?

- Abundance of web data useful for social scientists
 - social media
 - news media
 - government publications
 - organizational records
- Two kinds of ways to get data off the web
 - Web scraping: pulling information from user-facing websites for humans
 - APIs: application-facing structured information access, for computers

Web scraping vs. APIs

- Web scraping benefits
 - Any content that can be viewed on a website can be scraped
 - No API needed
 - No rate-limiting or authentication (usually)
- Web scraping challenges
 - Rarely tailored for information accessibility/organization
 - Your IP can be blocked
 - Messy, unstructured, inconsistent
 - Entirely site-dependent
- Rule of thumb: Check for API first. If not available, scrape.

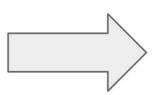
Some disclaimers

- Check a site's terms and conditions before scraping.
- Be nice don't hammer the site's server.
 - Add a time delay in between batches of scraping.
- Sites change their layout all the time. Your scraper will break

What is a website?

- Some combination of codebase, database.
- The "front end" product is HTML + CSS stylesheets + javascript.
- Browser turns the left image into the right.

hetail"Ha heed-"decator@ills.asp?MonterD-1218"HBills/AH/footh-/Ad-Hid width-"UA" hypolor="white" align="center" class="detail >=1200">Committees</are/font=</td>
*idth="150" bgoolor="white" align="center" clean="detail">550/font=
idth="150" bgoolor="center" clean="deta hetail"she/funter/admi/teretered width="804" homior="88895" class="datail"sta hesfo"/secate/Secator, aspidat9948ssbar(D-2794"she the"119" handlere"EMESS" allow"coster" elapse data! I bds house Washing Hell 19 Chillas (gird fonte /td) widthe" busher" of more "detail "Frie Arefe "Section | Nicos and Thomas | Nicos | Nico senate/Satutor_asgNike994Mamber(SedY)&"sJannifer Bartino-Sarrants/arn/Sontes/address widthe 184" bgcolore'white aligne'canter' cli materbills.asp@makerD=2005"=billst/g>-t/bett-t/id-tid widthe 10% homise-"white" aligne "center" class="detail"-to head-"denies 0-1216"+Committees=/aps/font=/bd=std width="154" hypologe"shine" align="center" class="datail">489-/font=-/bd=std width="154" hypologe"shine align="center" class="datail">489-/font=-/bd=std width="154" hypologe"shine hotall "she/footse/tds/trestrated widths "404" hopology "EMEMS" class datail "sea hreft" /anant/becator.asplide (Satumber (Sel) (Se 154" homine "ENERGY aligno" center" class "detail" or a heaft" fanatorbille, asp "Manher [8-2104" Billst (art) fonter/idend widths "154 moter class detail "He head demices than application of the light desired and the control of the second of the sec Setail "NEW/footed/advend widthe "15" Species "EMESS" aligns "center" class "detail "NEW/footed/adve/toversored widthe "65% Species smans/Sension_asg/Nds-99aMember[Dol]46"2fim Bisins/\anc/fontsc/td-rid widths 154" bgoslors white aligns center classs datail "rea odidi's@illar/abs/funtos/admind widths'ilh' bgooless'white' aligns'center' classs'detail'ssa heafs'deniumittees.asg/Wember20070 the "154" begoins "white" align="center" class="detail" HES/Font--/Ed-HES width="154" begoins "white" align="center" class="detail 104" homeles-"EMEME" class-"detail" his host-"/smate/Smator.asg/GA-MaRador/D-2108" William E. Bradys/arc/foot-c/tdrend width-onCommittees.aug/Months:25-1118 "eCommittees(/s)*(/tont)*(/ddftd uidth="154" hypology"EBERES align="center" class="dstail">644/font) "EMBER" align="center" slame="detail">Be/foot>=/td>=t/tr>=tor=td width="494" bgcolor="white" slame="detail">ta href="/secate/beca ader[5-1511"=Wallada Bushs/ars/fontrs/tdettd width="154" bgcolor="ehite" slign="center" class="detail"=ta href="SenatorBilla.asp! Widered width "134" bgmice "white" align "center" class "detail" his braft factomittees, as Phober (5-2211 "Specitess Cart/Enty "white" aligne conten" classe "datail" blic/fontec/adend widthe "lik" hymnion "white" aligne "conten" classe "datail "dec/fontec/adend "EMEMN" slame-"detail"-ta hosf-"/sensta/Benstor.aspND-994Monies25-2097"-Junes F. Clayborne, Dr. 4/a-4/font-4/detail width-"154" mentar' classe detail "sea brade denatorbilla aspinament 5-1017 "Abillar/ant/dustre/tdent widthe "154" bgooler="20100" align="cont. onformittees.sep?NamberUP=1097"nformitteest/abr/fidoritd.widtho"154" hopolor="EMEMER" aligno"center" class="detail"n\$34/fort) "THREES" align="center" class="detail"=80/fnoti=1/td=4/tr=tr=td=480 width="480" bgcolor="white" class="detail"=4a host=1/anata/Ban ober13-7133"-Jangueline V. Culline//a-c/fost-c/td-ftd width-"154" bgoolor-"ebita" align-"center" class-"detail"-ta bref-"denator wills "Millar/ans/fooths/tdstnd width="IN" bgoolor="white" align="center" class="detail"=ra href="beniumsittess.asg/Moones@hell the "154" homelors "white align="center" class="detail" #584/fonthe/bdf=fid width="154" homelors "white align="center" class="detail" HOAT Squalar #TERREST class="detail" ris hreft" /accate/Scrater_cap/NA-FORModer TD-2212" #Michael Connellyt/ar-(/Sont)-t/Advise visitormotor" class="detail">4a href="ImmatorBills.asg/WooderTD=1011">Bills=/an=/foot>=/foot>=/id>+td width="154" bgcolor="EMRMES" align="conto ocumnitrans.aug/Wooker33=1312"eCummitrans=/Jair/footh=/Job/ bycolor="EMESES" align="center" class="detail">21*/footh ""HMENN" aligne"center" classe"detail"=No/foot|=c/td=t/tr=tr=td="deta"| bpcslore"shipe" classe"detail"=na href="/secata/fens nder(D-2004"-Zuhn J. Callerton/Jahr/font-Clust-Club-Cid width-"154" hymior-"white" align-"center" class-"detail"-ta href-"denator-Wills t/bd>tid width: "IA" bgsclor="white" align="center" class="detall">ts bref="fenCommittees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016">Committees.asp?Monter:10-2016 r"white" align="center" class="detail"=6=/forces/odrend width="lik" bpoolor="white" align="center" class="detail"=6=/forces/odrend width='lik" bpoolor="white" align="center"=6=/forces/odrend width='lik"=6=/forces/odrend width='lik"=6=/ " "SMESSE" class="datail"==a href="/assate/Senator.aspNik=9taNesder:S=1213"=Thomas Gullertoon/abs/fontb=/fd="td" bpoulor



Leadership Officers Senate Seating Chart Democrats: 39 Republicans: 20					
Senator	Bills	Committees	District	Party	
Pamela J. Althoff	Bills	Committees	32	R	
Neil Anderson	Bills	Committees	36	R	
Jason A. Barickman	Bills	Committees	53	R	
Scott M. Bennett	Bills	Committees	52	D	
Jennifer Bertino-Tarrant	Bills	Committees	49	D	
Daniel Biss	Bills	Committees	9	D	
Tim Bivins	Bills	Committees	45	R	
William E. Brady	Bills	Committees	44	R	
Melinda Bush	Bills	Committees	31	D	
James F. Clayborne, Jr.	Bills	Committees	57	D	
Jacqueline Y. Collins	Bills	Committees	16	D	
Michael Connelly	Bills	Committees	21	R	
John J. Cullerton	Bills	Committees	6	D	

Current Senate Members 99th General Assembly

Web scraping returns HTML

- It's easy to pull HTML from a website
- It's much more difficult to find the information you want from that HTML.
- So we have to learn how to parse HTML to find the data we want

```
<html lang="en-us" class="a-js a-audio a-video a-canvas a-svg a-drag-drop a-geolocation a-history a-webworker a-autofocus a-input-placeh</pre>
older a-textarea-placeholder a-local-storage a-gradients a-hires a-transform3d -scrolling a-text-shadow a-text-stroke a-box-shadow a-borde
r-radius a-border-image a-opacity a-transform a-transition a-ember" data-19ax5a9jf="dingo" data-aui-build-date="3.21.9-2022-01-05">
   <!-- sp:feature:head-start -->
  ▶ <head>...</head>
   <!-- sp:end-feature:head-close -->
   <!-- sp:feature:start-body -->
  ▼<body class="a-m-us a-aui_72554-c a-aui_accordion_a11y_role_354025-c a-aui_killswitch_csa_logger_372963-c a-aui_launch_2021_ally_fixes
  _392482-c a-aui_pci_risk_banner_210084-c a-aui_preload_261698-c a-aui_rel_noreferrer_noopener_309527-c a-aui_template_weblab_cache_33340
  6-c a-aui tnr v2 180836-c a-meter-animate" data-new-gr-c-s-check-loaded="14.1043.0" data-gr-ext-installed>
   ▼<div id="a-page">
     ><script type="a-state" data-a-state="{"key":"a-wlab-states"}">...</script>
       <script>typeof uex === 'function' && uex('ld', 'portal-bb', {wb: 1})</script>
       <!-- sp:end-feature:start-body -->
     ▶ <script>...</script>
       <script>window.ue && ue.count && ue.count('CSMLibrarySize', 13275)</script>
       <!-- sp:feature:nav-inline-js -->
       <!-- NAVYAAN JS -->
     ▶ <script type="text/javascript">...</script>
     ▶ <script type="text/javascript">...</script>
       <img src="https://images-na.ssl-images-amazon.com/images/G/01/gno/sprites/nav-sprite-global-1x-hm-dsk-reorg. CB405937547 .png'</pre>
       style="display:none" alt>
       <script type="text/javascript">var nav t after preload sprite = + new Date();</script>
     ▶ <script>...</script>
       <!-- sp:end-feature:nav-inline-js -->
       <!-- sp:feature:nav-skeleton -->
       <!-- sp:end-feature:nav-skeleton -->
       <!-- sp:feature:navbar -->
       <!--Pilu -->
       <!-- NAVYAAN -->
       <!-- navmet initial definition -->
     ▶ <script type="text/javascript">...</script>
       <script type="text/javascript">window.navmet.tmp=+new Date();</script>
     ▶ <script type="text/javascript">...</script>
       <style mark="aboveNavInjectionCSS" type="text/css"> div#navSwmHoliday.nav-focus {border: none;margin: 0;} </style>
     ><script mark="aboveNavInjectionJS" type="text/javascript">...</script>
     ▶ <noscript>...</noscript>
       <script type="text/javascript">window.navmet.push({key:'PreNav',end:+new Date(),begin:window.navmet.tmp});</script>
       <a id="nav-top"></a>
       <a id="skiplink" tabindex="0" class="skip-link">Skip to main content</a>
       <script type="text/javascript">window.navmet.tmp=+new Date();</script>
       <!-- Navvaan Upnav -->
     ▶ <div id="nav-upnav" aria-hidden="true">...</div>
       <script type="text/javascript">window.navmet.push({key:'UpNav',end:+new Date(),begin:window.navmet.tmp});</script>
       <script type="text/javascript">window.navmet.main=+new Date();</script>
```

Basic strategy of web scraping

- Find out what kind of HTML element your data is in. Use your browser's "inspector".
- Think about how you can differentiate those elements from other, similar elements in the webpage using HTML/CSS anatomy.
 - This requires some basic knowledge of HTML/CSS.
 - We will go over some basic concepts below; just enough to get your started.
- Use Python and add-on modules like BeautifulSoup to extract just that data.

HTML: Basic structure

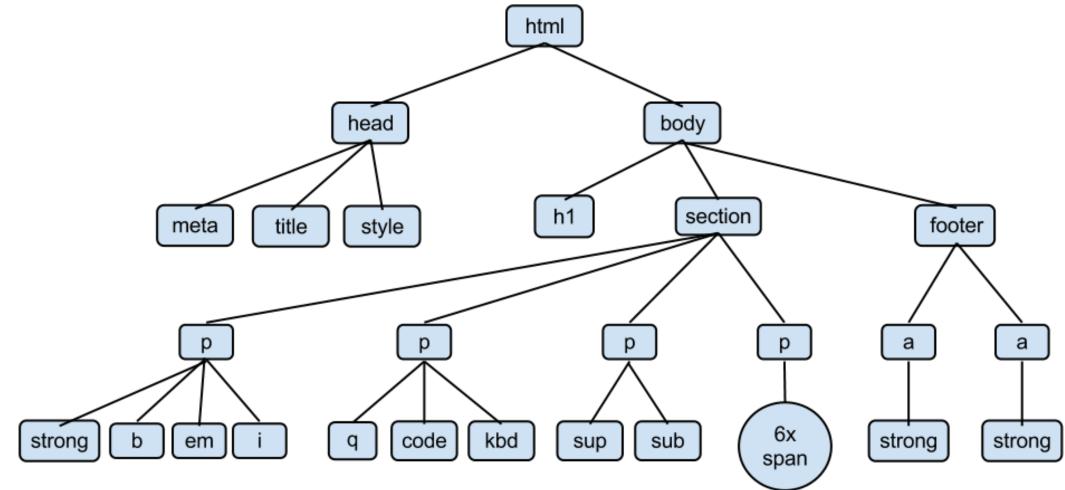
```
<!DOCTYPE html>
<html>
     <head>
       <title>Page title</title>
   </head>
    <body>
        Hello world!
    </body>
</html>
```

HTML: Basic structure

- The **head** contains metadata; data about the website. Things like the document title, styles, scripts.
- Examples:
 - <title>: The title of the document, which is required in all HTML/XHTML documents and is shown in the browser's title bar or on the page's tab.
 - <style>: CSS styles that are included directly within the HTML document.
 - <script>: Used to include JavaScript or link to external JavaScript files.
- The **body** element contains the content of an HTML document, such as text, images, links, tables, lists, etc. This is the section that is visible to the user in the web browser.
- Examples:
 - <h1> to <h6>: Header tags that define headings.
 - : Defines a paragraph.
 - <a>: Defines a hyperlink.
 - : Embeds an image.

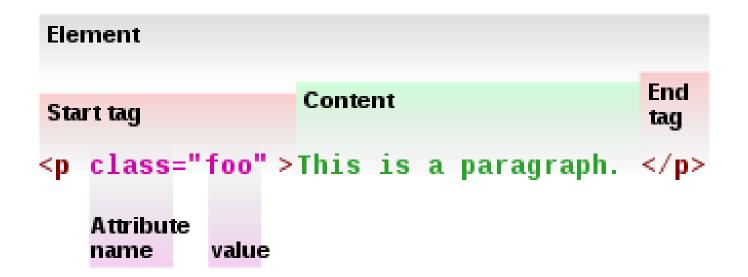
HTML as a tree

• Each branch of the tree is called an element

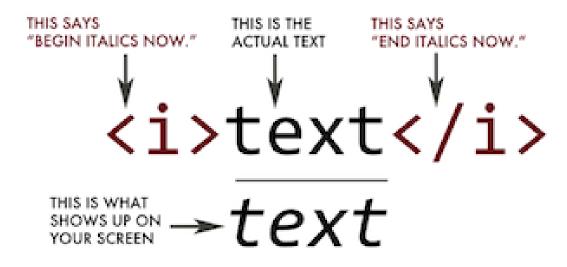


Three general components of HTML elements

- Tags (starting and ending the element)
- Attributes (giving information about the element)
- Text, or Content (the text inside the element)



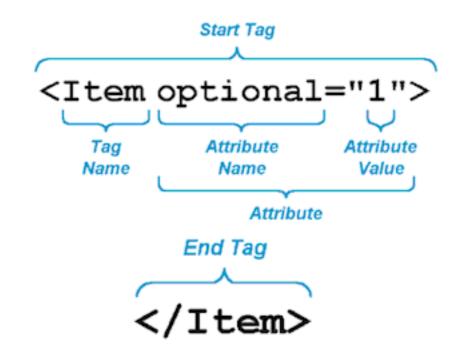
HTML tags



Tag	Meaning	
<head></head>	page header (metadata, etc	
<body></body>	holds all of the content	
	regular text (paragraph)	
<h1>,<h2>,<h3></h3></h2></h1>	header text, levels 1, 2, 3	
,,	ordered list, unordered list, list item	
</a 	link to "page.html"	
,,	table, table row, table item	
<div>,</div>	general containers (can contain CSS, JavaScript, etc.)	

HTML attributes

- HTML elements can have attributes.
- Attributes provide additional information about an element.
- Attributes are always specified in the start tag.
- Attributes come in name/value pairs like: name="value"



Finding data in HTML

- Sometimes we can find the data we want just by using HTML tags or attributes (e.g., all the <a> tags)
- More often, this isn't enough: There might be 1000 <a> tags on a page. But maybe we want only the <a> tags inside of a tag.
- This is where CSS comes in.

CSS (Cascading Style Sheet)

- CSS defines how HTML elements are to be displayed.
- HTML came first. But it was only meant to define content, not format it.
- CSS was created to display content on a webpage. Now, one can change the look of an entire website just by changing one file.
- Most web designers litter the HTML markup with tons of classes and ids to provide "hooks" for their CSS.
- You can piggyback on these "hooks" to jump to the parts of the HTML markup that contain the data you need.

CSS Selectors & Declarations

• Selector: a

• Property: background-color

• Value: blue

Туре	HTML	CSS Selector
Element	<a>>	a p a
Class		.blue a.blue
ID		#blue a#blue