

Week 1 Tutorial:

Web Scraping

POP77032 Quantitative Text Analysis for Social Scientists

Webpage Elements & Source Code

- Modern browsers come with very powerful built-in tools that help in web scraping
- Two most useful for us are:
 - Element inspection (Move cursor to some element on the page, Right click, Pick 'Inspect' from a menu)
 - Viewing page source (`Ctrl + U`/`Option + CMD + U` in Windows(Linux)/Mac for Chrome and Firefox)
- Knowing the source code and element attributes helps build XPath for selection

Element Inspection (Firefox)

The image is a screenshot of a web browser displaying the Wikipedia page "Members of the 1st Dáil". A right-click context menu is open over the table, showing options like "Save Page As...", "View Page Source", and "Inspect". The table lists members by constituency, name, and party. The bottom of the image shows the Chrome DevTools interface with the HTML and CSS panels visible. The HTML panel shows the table structure, and the CSS panel shows the styles for the table and its cells. The right sidebar of DevTools shows the "Layout" tab with a box model diagram.

Element Inspection (Chrome)

[illegible]

Source Code Inspection

```
view-source:https://en.wikipedia.org/wiki/Members_of_the_1st_Dáil - Chromium

W Members of the 1st Dáil - x W view-source:https://en.wi x +

view-source:https://en.wikipedia.org/wiki/Members_of_the_1st_Dáil

Line wrap
1 <!DOCTYPE html>
2 <html class="client-nojs" lang="en" dir="ltr">
3 <head>
4 <meta charset="UTF-8"/>
5 <title>Members of the 1st Dáil - Wikipedia</title>
6 <script>document.documentElement.className="client-js";RLCONF={{"wgBreakFrames":false,"wgSeparatorTransformTable":["",""],"wgDigitTransformTable":["",""],"wgDefaultDateFormat":"dmy","wgMonthNames":["","January","February","March","April","M
7 "All Wikipedia articles written in Hiberno-English","Use dmy dates from July 2021","Articles containing Irish-language text","Members of the 1st Dáil","1st Dáil","Lists of Teachtaí Dála by term"],"wgPageContentLanguage":"en","wgPageContent
8 "wgULSPosition":"interlanguage","wgULSisCompactLinksEnabled":true,"wgWikibaseItemId":"Q6814305","wgGENewcomerTasksGuidanceEnabled":true,"wgGEAskQuestionEnabled":false,"wgELinkRecommendationsFrontendEnabled":false};RLSTATE={{"ext.globalCssJ
9 "ext.visualEditor.targetLoader","ext.eventLogging","ext.wikimediaEvents","ext.navigationTiming","ext.cx.eventlogging.campaigns","ext.centralNotice.geoIP","ext.centralNotice.startUp","ext.centralauth.centralautologin","ext.uls.compactlinks"
10 <script>(RLQ=window.RLQ||[]).push(function(){mw.loader.implement("user.options@lhzgi",function($,jQuery,require,module){/*@nomin*/mw.user.tokens.set({"patrolToken":"+\\","watchToken":"+\\","csrfToken":"+\\"});
11 });});</script>
12 <link rel="stylesheet" href="/w/load.php?lang=en&modules=ext.cite.styles%7Cext.uls.interlanguage%7Cext.visualEditor.desktopArticleTarget.noscript%7Cext.wikimediaBadges%7Cjquery.makeCollapsible.styles%7Cskins.vector.styles.legacy%7Cwiki
13 <script async="" src="/w/load.php?lang=en&modules=startup&only=scripts&raw=1&skin=vector"></script>
14 <meta name="ResourceLoaderDynamicStyles" content="" />
15 <link rel="stylesheet" href="/w/load.php?lang=en&modules=site.styles&only=styles&skin=vector"/>
16 <meta name="generator" content="MediaWiki 1.38.0-wmf.9"/>
17 <meta name="referrer" content="origin"/>
18 <meta name="referrer" content="origin-when-crossorigin"/>
19 <meta name="referrer" content="origin-when-cross-origin"/>
20 <meta name="format-detection" content="telephone=no"/>
21 <meta property="og:image" content="https://upload.wikimedia.org/wikipedia/en/5/57/Ist-meeting.jpg"/>
22 <meta property="og:image:width" content="1200"/>
23 <meta property="og:image:height" content="826"/>
24 <meta property="og:image" content="https://upload.wikimedia.org/wikipedia/en/thumb/5/57/Ist-meeting.jpg/800px-Ist-meeting.jpg"/>
25 <meta property="og:image:width" content="800"/>
26 <meta property="og:image:height" content="551"/>
27 <meta property="og:image" content="https://upload.wikimedia.org/wikipedia/en/thumb/5/57/Ist-meeting.jpg/640px-Ist-meeting.jpg"/>
28 <meta property="og:image:width" content="640"/>
29 <meta property="og:image:height" content="441"/>
30 <meta property="og:title" content="Members of the 1st Dáil - Wikipedia"/>
31 <meta property="og:type" content="website"/>
32 <link rel="preconnect" href="//upload.wikimedia.org"/>
33 <link rel="alternate" media="only screen and (max-width: 720px)" href="//en.m.wikipedia.org/wiki/Members_of_the_1st_D%C3%A1il"/>
34 <link rel="alternate" type="application/x-wiki" title="Edit this page" href="/w/index.php?title=Members_of_the_1st_D%C3%A1il&action=edit"/>
35 <link rel="apple-touch-icon" href="/static/apple-touch/wikipedia.png"/>
36 <link rel="shortcut icon" href="/static/favicon/wikipedia.ico"/>
37 <link rel="search" type="application/opensearchdescription+xml" href="/w/opensearch_desc.php" title="Wikipedia (en)"/>
38 <link rel="EditURI" type="application/rsd+xml" href="//en.wikipedia.org/w/api.php?action=rsd"/>
39 <link rel="license" href="https://creativecommons.org/licenses/by-sa/3.0/" />
40 <link rel="canonical" href="https://en.wikipedia.org/wiki/Members_of_the_1st_D%C3%A1il"/>
41 <link rel="dns-prefetch" href="//meta.wikimedia.org" />
42 <link rel="dns-prefetch" href="//login.wikimedia.org"/>
43 </head>
44 <body class="mediawiki ltr sitedir-ltr mw-hide-empty-elt ns-0 ns-subject mw-editable page-Members_of_the_1st_Dáil rootpage-Members_of_the_1st_Dáil skin-vector action-view skin-vector-legacy"><div id="mw-page-base" class="noprint"></div>
45 <div id="mw-head-base" class="noprint"></div>
46 <div id="content" class="mw-body" role="main">
47 <a id="top"></a>
48 <div id="siteNotice"><!-- CentralNotice --></div>
49 <div class="mw-indicators">
50 </div>
51 <h1 id="firstHeading" class="firstHeading">Members of the 1st Dáil</h1>
52 <div id="bodyContent" class="vector-body">
53 <div id="siteSub" class="noprint">From Wikipedia, the free encyclopedia</div>
54 <div id="contentSub"></div>
55 <div id="contentSub2"></div>
56
57 <div id="jump-to-nav"></div>
58 <a class="mw-jump-link" href="#mw-head">Jump to navigation</a>
59 <a class="mw-jump-link" href="#searchInput">Jump to search</a>
60 <div id="mw-content-text" class="mw-body-content mw-content-ltr" lang="en" dir="ltr"><div class="mw-parser-output"><div class="shortdescription nomobile noexcerpt noprint searchaux" style="display:none">List of TDs in the 1st Dáil<
61 <table class="box More citations needed plainlinks metadata ambbox ambbox-content ambbox-Refimprove" role="presentation"><tbody><tr><td class="mbx-image"><div style="width:52px"><a href="/wiki/File:Question_book_new.svg" class="image"><img a
```

Exercise 1: Working with HTML

- Consider a simple HTML code below.
- Read it into R using `rvest`'s `read_html()` function.
- How many children does the `<body>` element have?
- Build an XPath to extract the first paragraph.
- Build an XPath to extract the second sub-heading under a green header.

```
1 library("rvest")

1 html_txt <- "
2 <!DOCTYPE html>
3 <html>
4   <head>
5     <title>A title</title>
6   </head>
7   <body>
8     <h1 style='color:Red;'>
9       Heading 1
10    <h2>Subheading</h2>
11    </h1>
12    <h1 style='color:Green;'>
13      Heading 2
14      <h2>Subheading 1</h2>
15      <h2>Subheading 2</h2>
16    </h1>
17    <p>A paragraph.</p>
18    <p>Another paragraph.</p>
19  </body>
20 </html>"
```

Exercise 2: Working with Webpages

- Now let's turn to a real website.
- Here we will extract the table of countries with their GDP from a Wikipedia article
- Start by loading in the webpage using `rvest`'s `read_html()` function.
- Go the webpage of the article and locate the elements that would be helpful for table extraction.
- Extract the `<table>` node that corresponds to the main table.
- Extract `<tbody>` element as a child of this element
- Extract the table of with data using `rvest`'s `html_table()` function
- Tidy up the extracted table