Harvesting online data

Short course at Santa Clara University, ECON173

Johannes Fritz

5/5/22, 5/12/22

WHY AM I HERE?

- ► Economist with a strong “data science tilt”
  - > Web harvesting
  - > Machine learning
  - > Natural language processing

- ► Data collection is my day job
  - > Global Trade Alert
    Dataset on industrial policy choices around the world since 2009
  - > Digital Policy Alert
    Dataset on G20+Europe’s digital policy choices since 2021

YOU WILL LEARN 3 TOOLS

1. Harvesting the data:
   Headless **web scraping** using R.

today

2. Parsing the data:
   Introduction to **Regular Expressions**.

3. Organizing the data:
   Basics of **relational database design**.

next week

All material at:
https://github.com/johannesfritz/scu-harvesting

Getting started

- What projects do you want to tackle?
- What experience do you already have?
- What tools do you want to learn?

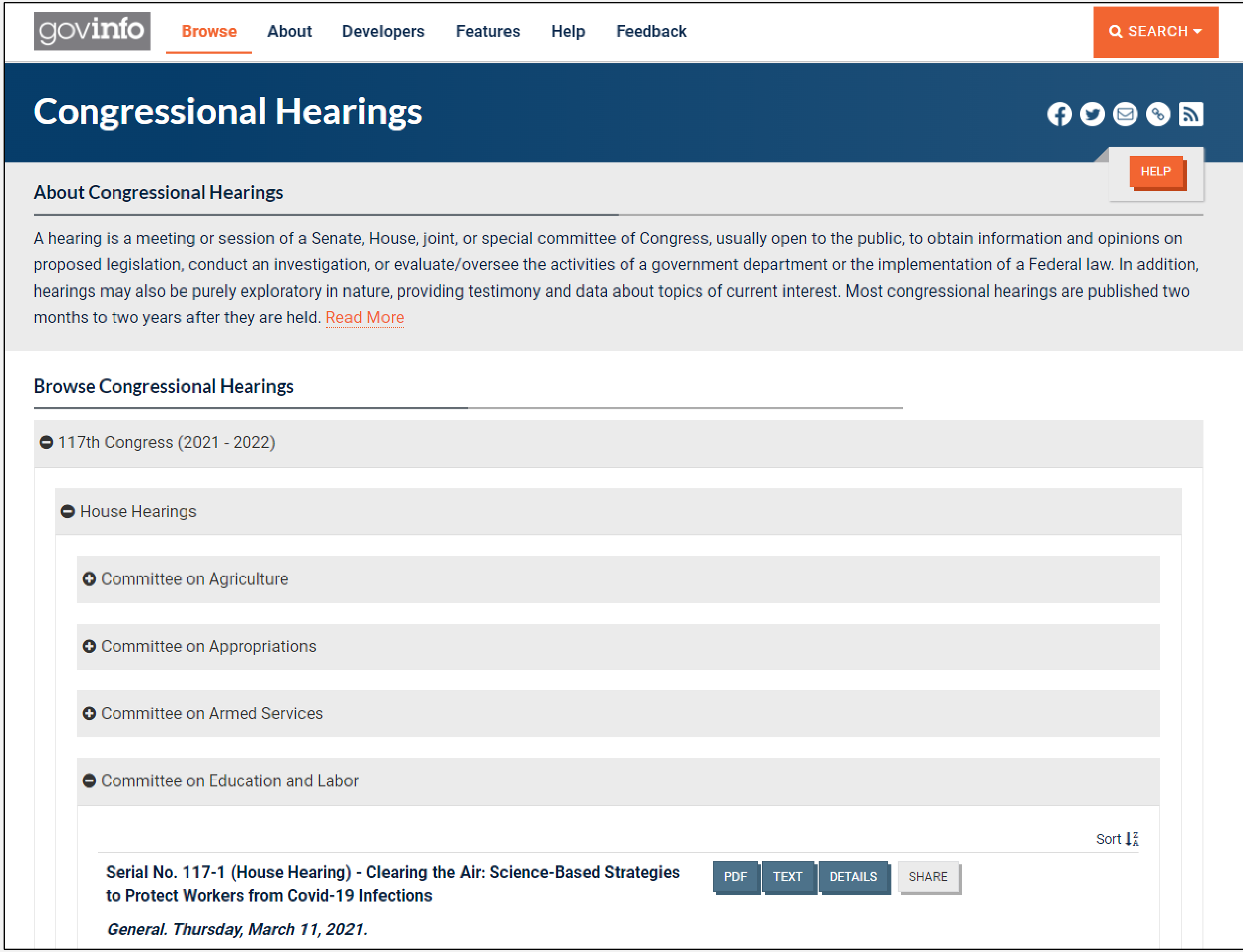
The U.S. Congressional Record

My teaching example for this short course

The U.S. Congressional Hearings

Source

- https://www.govinfo.gov/app/collection/CHRG/
- or search: «GPO Congressional Hearings»



Congressional Hearings

About Congressional Hearings

A hearing is a meeting or session of a Senate, House, joint, or special committee of Congress, usually open to the public, to obtain information and opinions on proposed legislation, conduct an investigation, or evaluate/oversee the activities of a government department or the implementation of a Federal law. In addition, hearings may also be purely exploratory in nature, providing testimony and data about topics of current interest. Most congressional hearings are published two months to two years after they are held. Read More

Browse Congressional Hearings

- 117th Congress (2021 - 2022)
  - House Hearings
    - Committee on Agriculture
    - Committee on Appropriations
    - Committee on Armed Services
    - Committee on Education and Labor

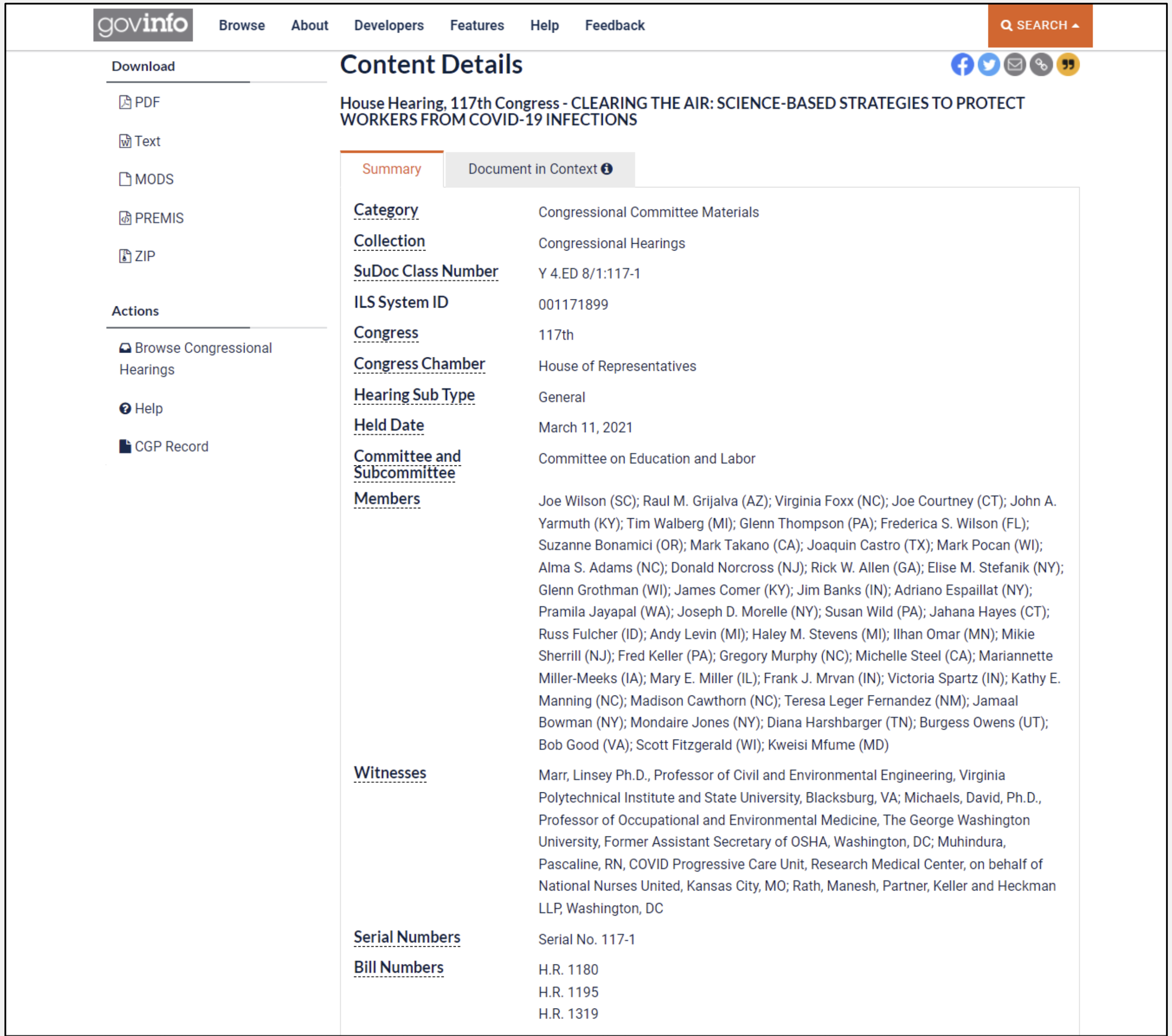
Sort

Serial No. 117-1 (House Hearing) - Clearing the Air: Science-Based Strategies to Protect Workers from Covid-19 Infections

PDF TEXT DETAILS SHARE

General. Thursday, March 11, 2021.

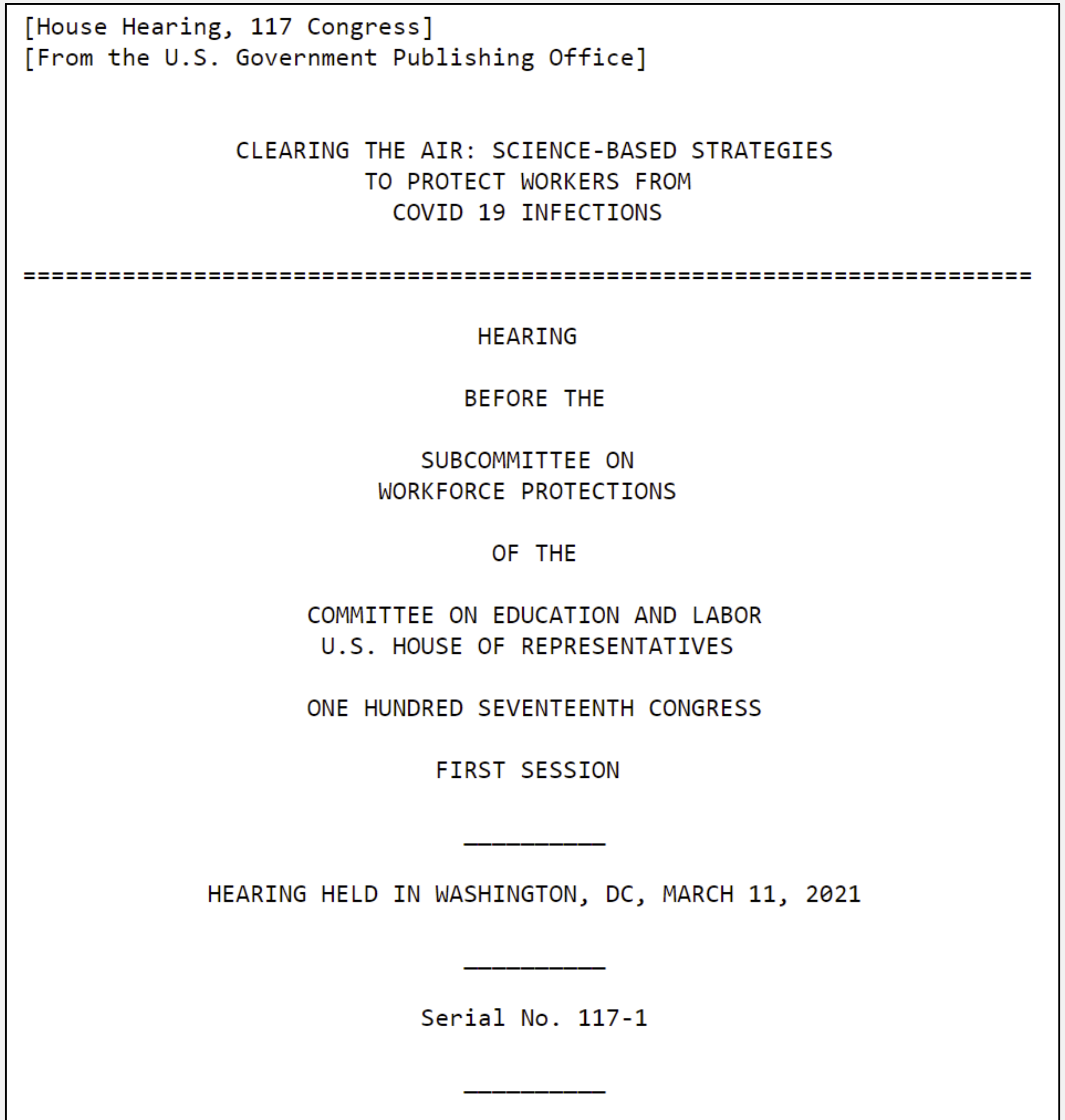
WEBSITE CONTENTS



Content Details

House Hearing, 117th Congress - CLEARING THE AIR: SCIENCE-BASED STRATEGIES TO PROTECT WORKERS FROM COVID-19 INFECTIONS

Category	Congressional Committee Materials
Collection	Congressional Hearings
SuDoc Class Number	Y 4.ED 8/1:117-1
ILS System ID	001171899
Congress	117th
Congress Chamber	House of Representatives
Hearing Sub Type	General
Held Date	March 11, 2021
Committee and Subcommittee	Committee on Education and Labor
Members	Joe Wilson (SC); Raul M. Grijalva (AZ); Virginia Foxx (NC); Joe Courtney (CT); John A. Yarmuth (KY); Tim Walberg (MI); Glenn Thompson (PA); Frederica S. Wilson (FL); Suzanne Bonamici (OR); Mark Takano (CA); Joaquin Castro (TX); Mark Pocan (WI); Alma S. Adams (NC); Donald Norcross (NJ); Rick W. Allen (GA); Elise M. Stefanik (NY); Glenn Grothman (WI); James Comer (KY); Jim Banks (IN); Adriano Espaillat (NY); Pramila Jayapal (WA); Joseph D. Morelle (NY); Susan Wild (PA); Jahana Hayes (CT); Russ Fulcher (ID); Andy Levin (MI); Haley M. Stevens (MI); Ilhan Omar (MN); Mikie Sherrill (NJ); Fred Keller (PA); Gregory Murphy (NC); Michelle Steel (CA); Mariannette Miller-Meeks (IA); Mary E. Miller (IL); Frank J. Mrvan (IN); Victoria Spartz (IN); Kathy E. Manning (NC); Madison Cawthorn (NC); Teresa Leger Fernandez (NM); Jamaal Bowman (NY); Mondaire Jones (NY); Diana Harshbarger (TN); Burgess Owens (UT); Bob Good (VA); Scott Fitzgerald (WI); Kweisi Mfume (MD)
Witnesses	Marr, Linsey Ph.D., Professor of Civil and Environmental Engineering, Virginia Polytechnical Institute and State University, Blacksburg, VA; Michaels, David, Ph.D., Professor of Occupational and Environmental Medicine, The George Washington University, Former Assistant Secretary of OSHA, Washington, DC; Muhindura, Pascaline, RN, COVID Progressive Care Unit, Research Medical Center, on behalf of National Nurses United, Kansas City, MO; Rath, Manesh, Partner, Keller and Heckman LLP, Washington, DC
Serial Numbers	Serial No. 117-1
Bill Numbers	H.R. 1180 H.R. 1195 H.R. 1319



```
[House Hearing, 117 Congress]
[From the U.S. Government Publishing Office]


                CLEARING THE AIR: SCIENCE-BASED STRATEGIES
                       TO PROTECT WORKERS FROM
                         COVID 19 INFECTIONS

=======================================================================

                               HEARING

                              BEFORE THE

                           SUBCOMMITTEE ON
                        WORKFORCE PROTECTIONS

                                OF THE

                   COMMITTEE ON EDUCATION AND LABOR
                    U.S. HOUSE OF REPRESENTATIVES

                   ONE HUNDRED SEVENTEENTH CONGRESS

                            FIRST SESSION

                              __________

             HEARING HELD IN WASHINGTON, DC, MARCH 11, 2021

                              __________

                           Serial No. 117-1

                              __________
```

BUT:
HEARINGS ARE NOT RESEARCH READY

No search, nor filtered export

- Dates, topics
- Members of Congress, Witnesses
- Bills
- Individual speeches

Have to create the dataset ourselves!

1. Harvest it.
2. Parse it.
3. Store it.

YOU WILL LEARN 3 TOOLS

1. Harvesting the data:
   Headless **web scraping** using R.

2. Parsing the data:
   Introduction to **Regular Expressions**.

3. Organizing the data:
   Basics of **relational database design**.

All material at:
https://github.com/johannesfritz/scu-harvesting

1.

Harvesting the data

Headless web scraping using R

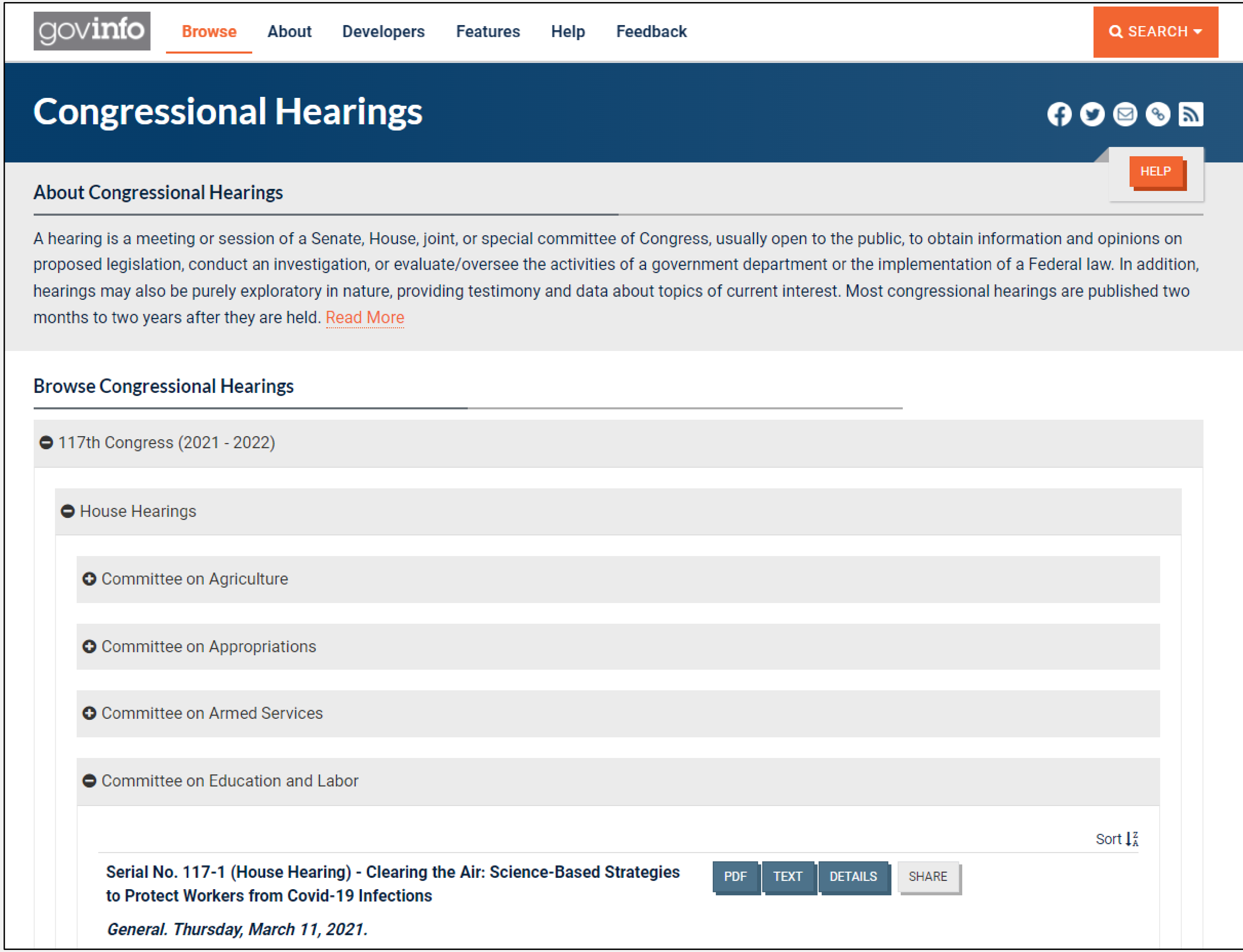
TWO APPROACHES TO HARVESTING WEBSITES

1. GET requests
   > Effectively downloading the HTML/etc. files
   > R library {httr}

2. Browser automation + GET requests
   > Used for website testing
   > «headless»: without open window (GUI)
   > R library {webdriver}

WHY WE NEED BROWSER AUTOMATION.

- Don’t know all the URLs.
- Site is dynamically rendered.



Congressional Hearings

About Congressional Hearings

A hearing is a meeting or session of a Senate, House, joint, or special committee of Congress, usually open to the public, to obtain information and opinions on proposed legislation, conduct an investigation, or evaluate/oversee the activities of a government department or the implementation of a Federal law. In addition, hearings may also be purely exploratory in nature, providing testimony and data about topics of current interest. Most congressional hearings are published two months to two years after they are held. Read More

Browse Congressional Hearings

- 117th Congress (2021 - 2022)
  - House Hearings
    - Committee on Agriculture
    - Committee on Appropriations
    - Committee on Armed Services
    - Committee on Education and Labor

Sort

Serial No. 117-1 (House Hearing) - Clearing the Air: Science-Based Strategies to Protect Workers from Covid-19 Infections

PDF TEXT DETAILS SHARE

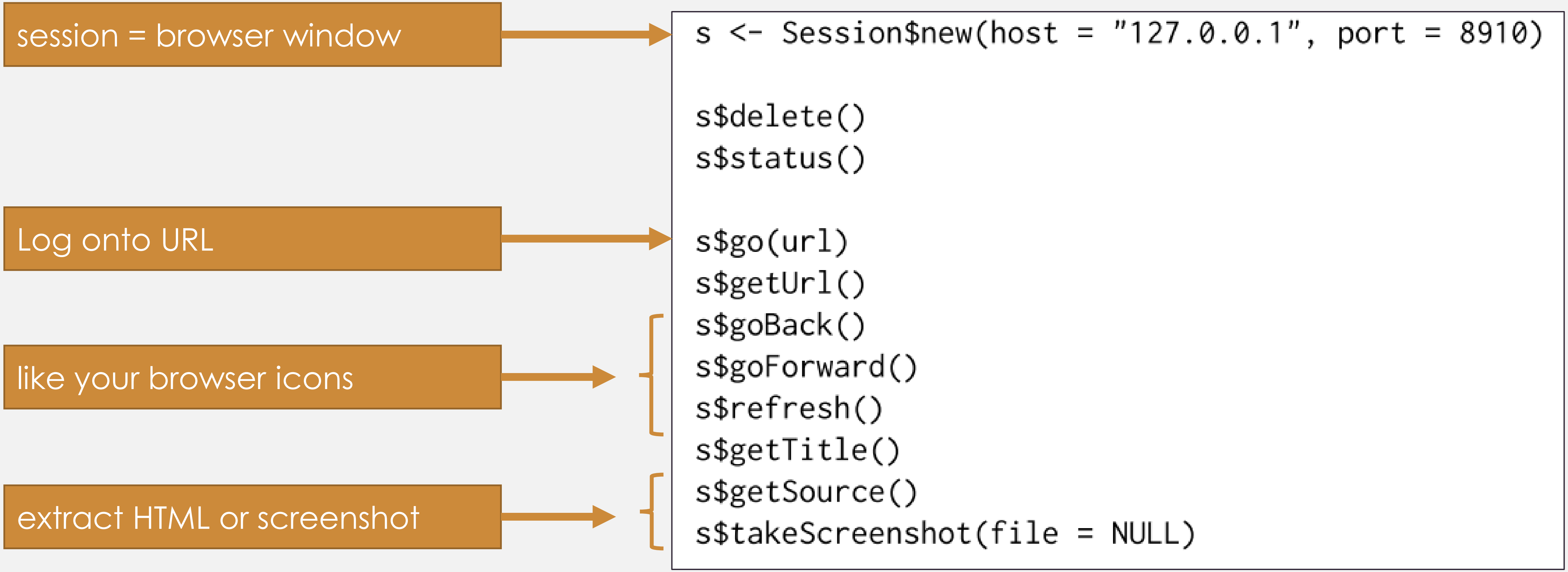
General. Thursday, March 11, 2021.

{WEBDRIVER} BASICS

- ► Library supporting «headless» browsing
  - > Using phantomJS
  - > Intuitive for rapid use
  - > Flexible extensions (if you know JavaScript)
  - > Alternative: {Rselenium}

- ► Supports
  - > Browsing
  - > Manipulating the page

{WEBDRIVER} BASICS
BROWSING



session = browser window

Log onto URL

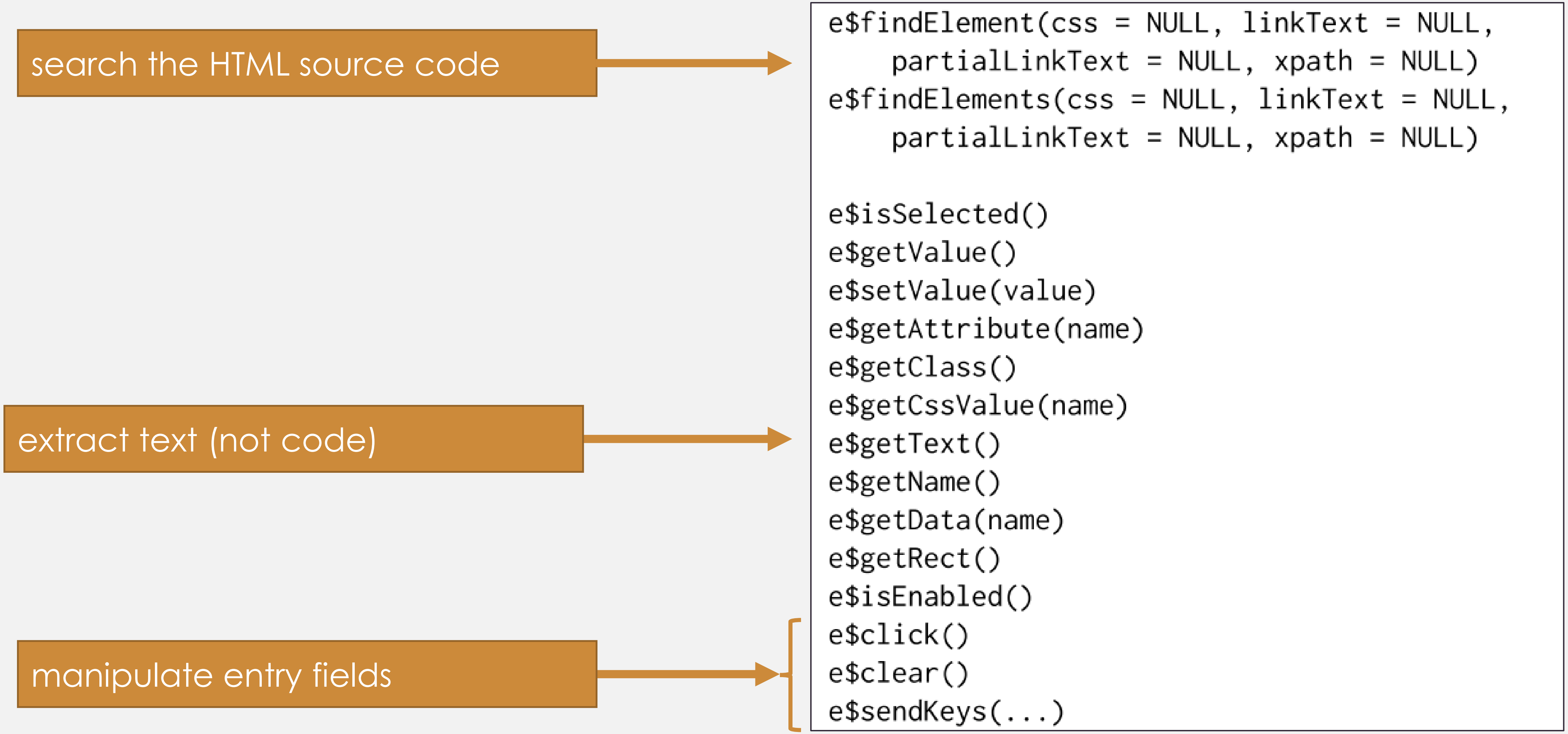
like your browser icons

extract HTML or screenshot

```
s <- Session$new(host = "127.0.0.1", port = 8910)

s$delete()
s$status()

s$go(url)
s$getUrl()
s$goBack()
s$goForward()
s$refresh()
s$getTitle()
s$getSource()
s$takeScreenshot(file = NULL)
```

{WEBDRIVER} BASICS

LOCATING + MANIPULATING TARGETS



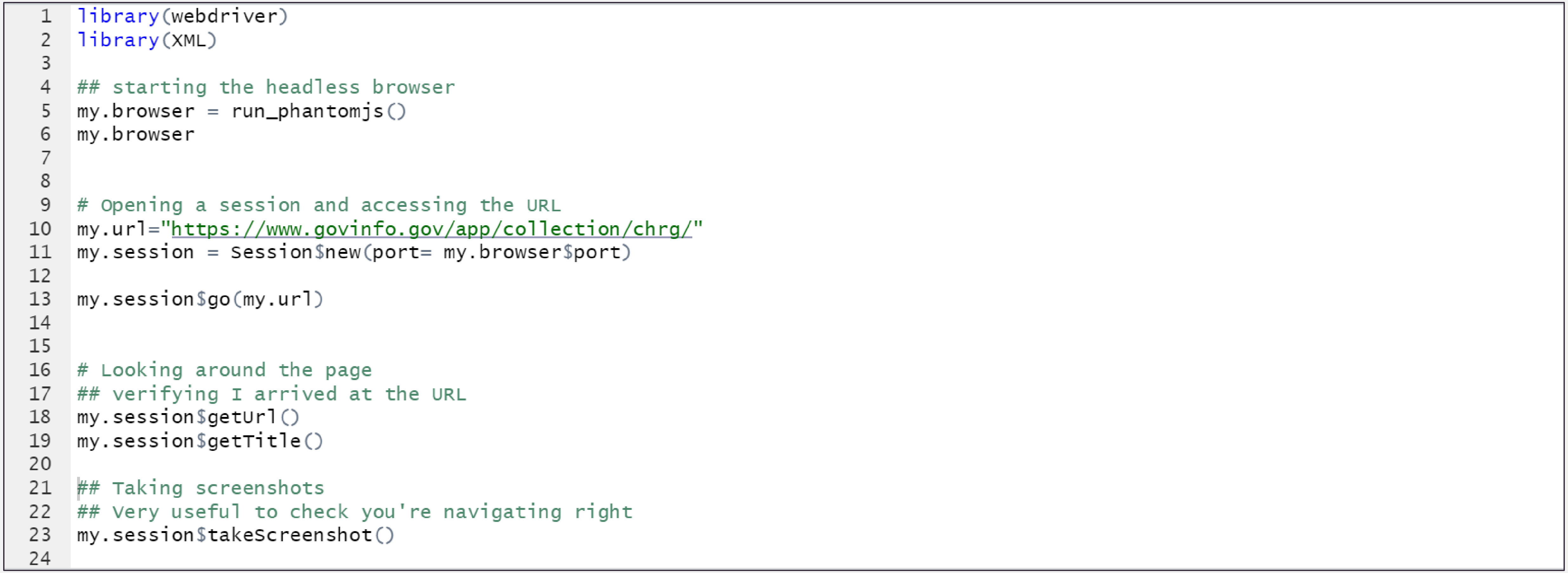
search the HTML source code

extract text (not code)

manipulate entry fields

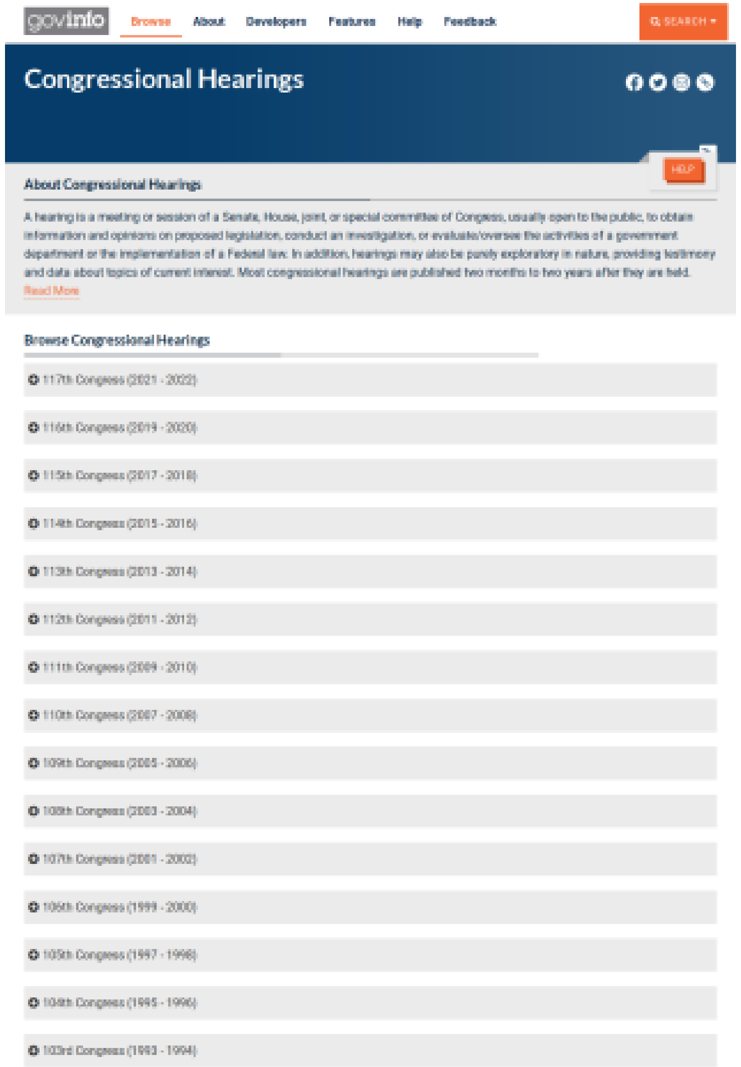
```
e$findElement(css = NULL, linkText = NULL,
    partialLinkText = NULL, xpath = NULL)
e$findElements(css = NULL, linkText = NULL,
    partialLinkText = NULL, xpath = NULL)

e$isSelected()
e$getValue()
e$setValue(value)
e$getAttribute(name)
e$getClass()
e$getCssValue(name)
e$getText()
e$getName()
e$getData(name)
e$getRect()
e$isEnabled()
e$click()
e$clear()
e$sendKeys(...)
```

Our initial script



```
library(webdriver)
library(XML)

## starting the headless browser
my.browser = run_phantomjs()
my.browser


# Opening a session and accessing the URL
my.url="https://www.govinfo.gov/app/collection/chrg/"
my.session = Session$new(port= my.browser$port)

my.session$go(my.url)


# Looking around the page
## verifying I arrived at the URL
my.session$getUrl()
my.session$getTitle()

## Taking screenshots
## Very useful to check you're navigating right
my.session$takeScreenshot()

```

Super useful: Taking screenshots



Congressional Hearings

About Congressional Hearings

A hearing is a meeting or session of a Senate, House, joint, or special committee of Congress, usually open to the public, to obtain information and opinions on proposed legislation, conduct an investigation, or evaluate/oversee the activities of a government department or the implementation of a Federal law. In addition, hearings may also be purely exploratory in nature, providing testimony and data about topics of current interest. Most congressional hearings are published two months to two years after they are held.

Read More

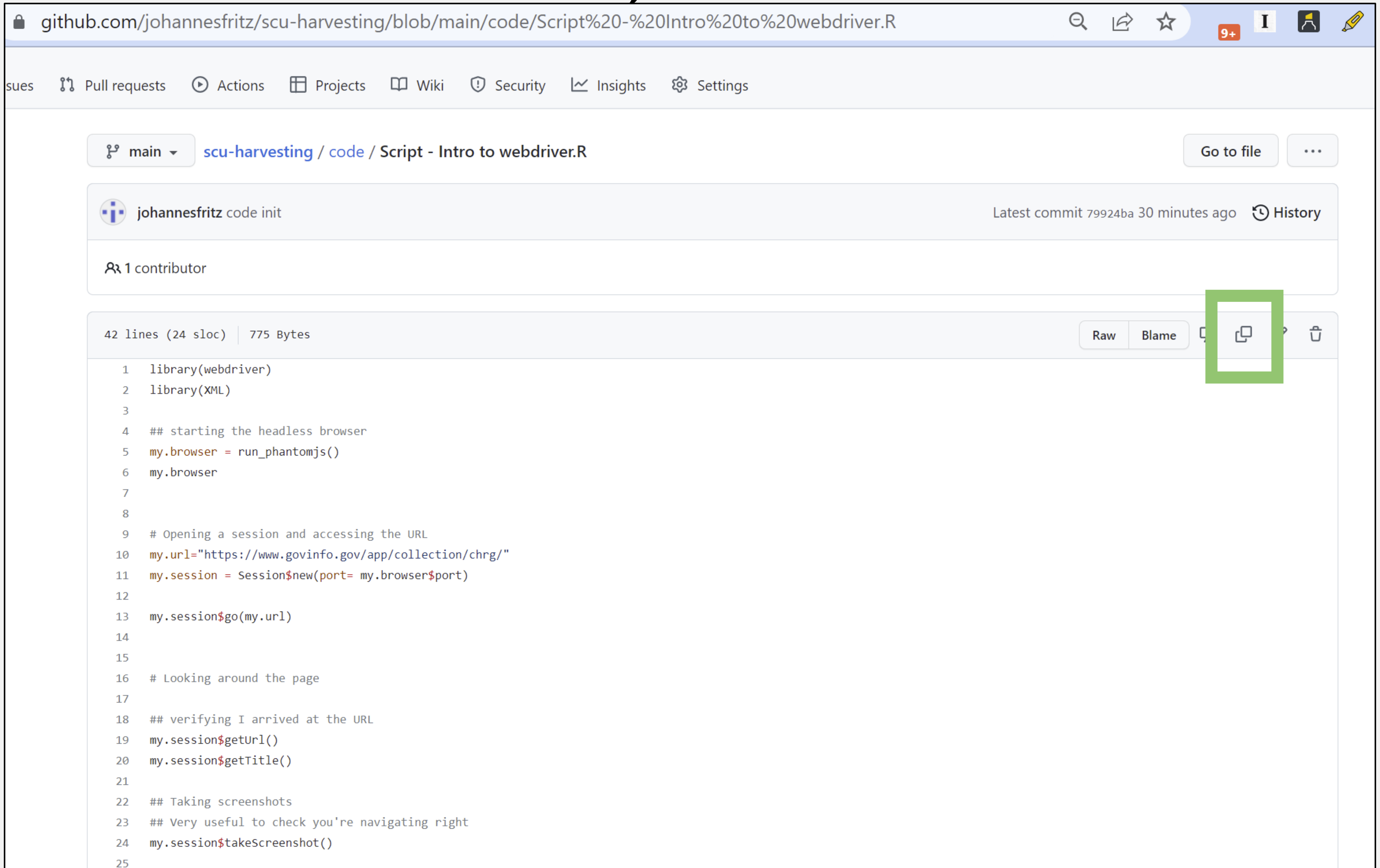
Browse Congressional Hearings

- 117th Congress (2021 - 2022)
- 116th Congress (2019 - 2020)
- 115th Congress (2017 - 2018)
- 114th Congress (2015 - 2016)
- 113th Congress (2013 - 2014)
- 112th Congress (2011 - 2012)
- 111th Congress (2009 - 2010)
- 110th Congress (2007 - 2008)
- 109th Congress (2005 - 2006)
- 108th Congress (2003 - 2004)
- 107th Congress (2001 - 2002)
- 106th Congress (1999 - 2000)
- 105th Congress (1997 - 1998)
- 104th Congress (1995 - 1996)
- 103rd Congress (1993 - 1994)

LOAD THE SCRIPT + SEE IF IT WORKS.

TRY OUT YOUR OWN PAGE, TOO

Copy it from GitHub



github.com/johannesfritz/scu-harvesting/blob/main/code/Script%20-%20Intro%20to%20webdriver.R

scu-harvesting / code / Script - Intro to webdriver.R

johannesfritz code init

Latest commit 79924ba 30 minutes ago

1 contributor

42 lines (24 sloc) | 775 Bytes

```
library(webdriver)
library(XML)

## starting the headless browser
my.browser = run_phantomjs()
my.browser


# Opening a session and accessing the URL
my.url="https://www.govinfo.gov/app/collection/chrg/"
my.session = Session$new(port= my.browser$port)

my.session$go(my.url)


# Looking around the page

## verifying I arrived at the URL
my.session$getUrl()
my.session$getTitle()

## Taking screenshots
## Very useful to check you're navigating right
my.session$takeScreenshot()
```

MOVING ON:
KEY ELEMENTS OF OUR FULL SCRIPT

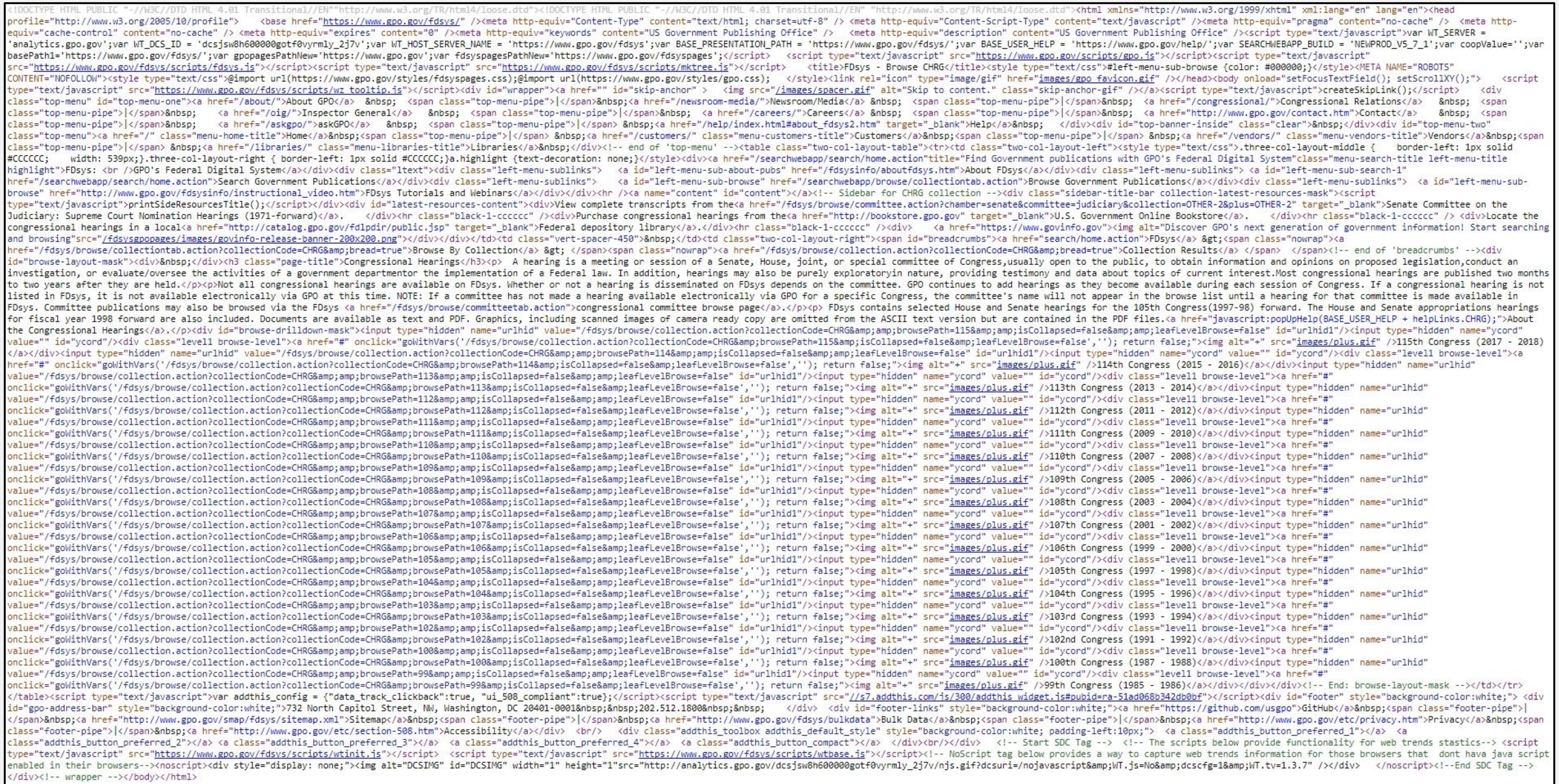
1. Navigate the site:
   need to tell our browser where to click
   > Tool: XPath

2. Create our local copy:
   store all pages of interest

3. Make capture comprehensive:
   go through all the pages we need
   > Tool: for loops

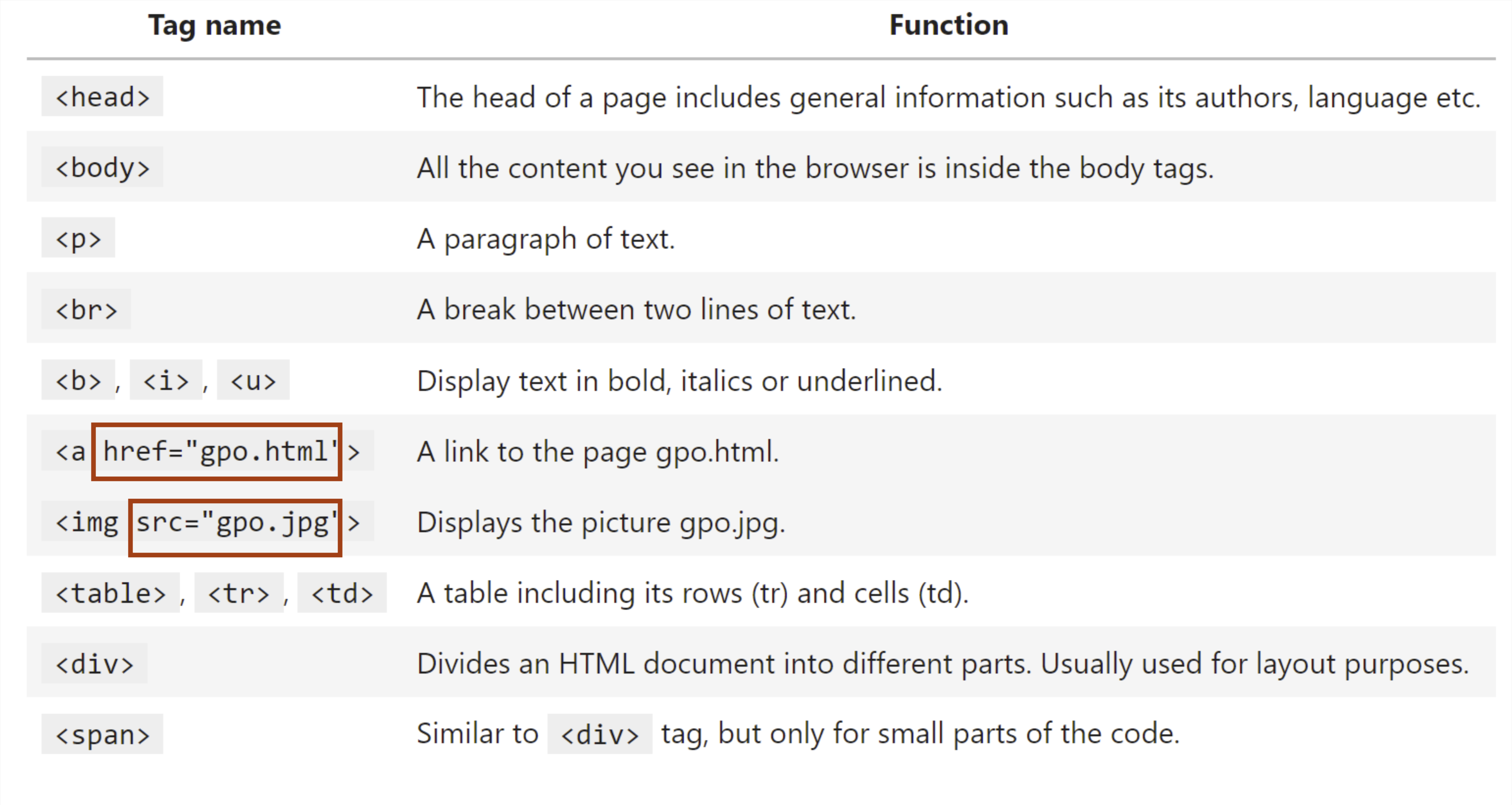
1. Navigate the site

the Document Object model (DOM)



```
<!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN""http://www.w3.org/TR/html4/loose.dtd"><!DOCTYPE HTML PUBLIC "-//W3C//DTD HTML 4.01 Transitional//EN" "http://www.w3.org/TR/html4/loose.dtd"><html xmlns="http://www.w3.org/1999/xhtml" xml:lang="en" lang="en"><head
profile="http://www.w3.org/2005/10/profile">    <base href="https://www.gpo.gov/fdsys/" /><meta http-equiv="Content-Type" content="text/html; charset=utf-8" /> <meta http-equiv="Content-Script-Type" content="text/javascript" /><meta http-equiv="pragma" content="no-cache" />  <meta http-
equiv="cache-control" content="no-cache" /> <meta http-equiv="expires" content="0" /><meta http-equiv="keywords" content="US Government Publishing Office" />   <meta http-equiv="description" content="US Government Publishing Office" /><script type="text/javascript">var WT_SERVER =
'analytics.gpo.gov';var WT_DCS_ID = 'dcsjsw8h600000gotf0vyrmly_2j7v';var WT_HOST_SERVER_NAME = 'https://www.gpo.gov/fdsys';var BASE_PRESENTATION_PATH = 'https://www.gpo.gov/fdsys/';var BASE_USER_HELP = 'https://www.gpo.gov/help/';var SEARCHWEBAPP_BUILD = 'NEWPROD_V5_7_1';var coopValue='';var
basePath1='https://www.gpo.gov/fdsys/';var gpopagesPathNew='https://www.gpo.gov';var fdsyspagesPathNew='https://www.gpo.gov/fdsyspages';</script>   <script type="text/javascript" src="https://www.gpo.gov/scripts/gpo.js"></script><script type="text/javascript"
src="https://www.gpo.gov/fdsys/scripts/fdsys.js"></script><script type="text/javascript" src="https://www.gpo.gov/fdsys/scripts/mktree.js"></script>    <title>FDsys - Browse CHRG</title><style type="text/css">#left-menu-sub-browse {color: #000000;}</style><META NAME="ROBOTS"
CONTENT="NOFOLLOW"><style type="text/css">@import url(https://www.gpo.gov/styles/fdsyspages.css);@import url(https://www.gpo.gov/styles/gpo.css);   </style><link rel="icon" type="image/gif" href="images/gpo_favicon.gif" /></head><body onload="setFocusTextField(); setScrollXY();">    <script
type="text/javascript" src="https://www.gpo.gov/fdsys/scripts/wz_tooltip.js"></script><div id="wrapper"><a href="" id="skip-anchor" >   <img src="/images/spacer.gif" alt="Skip to content." class="skip-anchor-gif" /></a><script type="text/javascript">createSkipLink();</script>    <div
class="top-menu" id="top-menu-one"><a href="/about/">About GPO</a>  &nbsp;  <span class="top-menu-pipe">|</span>&nbsp;<a href="/newsroom-media/">Newsroom/Media</a> &nbsp;  <span class="top-menu-pipe">|</span>&nbsp;  <a href="/congressional/">Congressional Relations</a>   &nbsp;  <span
class="top-menu-pipe">|</span>&nbsp;    <a href="/oig/">Inspector General</a>   &nbsp;  <span class="top-menu-pipe">|</span>&nbsp;  <a href="/careers/">Careers</a> &nbsp;  <span class="top-menu-pipe">|</span>&nbsp;  <a href="http://www.gpo.gov/contact.htm">Contact</a>    &nbsp;  <span
class="top-menu-pipe">|</span>&nbsp;    <a href="/askgpo/">askGPO</a>   &nbsp;  <span class="top-menu-pipe">|</span> &nbsp;<a href="/help/index.html#about_fdsys2.htm" target="_blank">Help</a>&nbsp;   </div><div id="top-banner-inside" class="clear">&nbsp;</div><div id="top-menu-two"
class="top-menu"><a href="/" class="menu-home-title">Home</a>&nbsp;<span class="top-menu-pipe">|</span> &nbsp;<a href="/customers/" class="menu-customers-title">Customers</a>&nbsp;<span class="top-menu-pipe">|</span> &nbsp;<a href="/vendors/" class="menu-vendors-title">Vendors</a>&nbsp;<span
class="top-menu-pipe">|</span> &nbsp;<a href="/libraries/" class="menu-libraries-title">Libraries</a>&nbsp;</div><!-- end of 'top-menu' --><table class="two-col-layout-table"><tr><td class="two-col-layout-left"><style type="text/css">.three-col-layout-middle {    border-left: 1px solid
#CCCCCC;    width: 539px;}.three-col-layout-right { border-left: 1px solid #CCCCCC;}a.highlight {text-decoration: none;}</style><div><a href="/searchwebapp/search/home.action"title="Find Government publications with GPO's Federal Digital System"class="menu-search-title left-menu-title
highlight">FDsys: <br />GPO's Federal Digital System</a></div><div class="ltext"><div class="left-menu-sublinks">   <a id="left-menu-sub-about-pubs" href="/fdsysinfo/aboutfdsys.htm">About FDsys</a></div><div class="left-menu-sublinks"> <a id="left-menu-sub-search-1"
href="/searchwebapp/search/home.action">Search Government Publications</a></div><div class="left-menu-sublinks">    <a id="left-menu-sub-browse" href="/searchwebapp/browse/collectiontab.action">Browse Government Publications</a></div><div class="left-menu-sublinks">  <a id="left-menu-sub-
browse" href="http://www.gpo.gov/fdsysinfo/instructional_video.htm">FDsys Tutorials and Webinars</a></div></div><hr /><a name="content" id="content"></a><!-- Sidebar for CHRG collection --><div class="sidebar-title-bar collection-latest-resources-mask"><script
type="text/javascript">printSideResourcesTitle();</script></div><div id="latest-resources-content"><div>View complete transcripts from the<a href="/fdsys/browse/committee.action?chamber=senate&committee=judiciary&collection=OTHER-2&plus=OTHER-2" target="_blank">Senate Committee on the
Judiciary: Supreme Court Nomination Hearings (1971-forward)</a>.    </div><hr class="black-1-cccccc" /><div>Purchase congressional hearings from the<a href="http://bookstore.gpo.gov" target="_blank">U.S. Government Online Bookstore</a>.    </div><hr class="black-1-cccccc" /> <div>Locate the
congressional hearings in a local<a href="http://catalog.gpo.gov/fdlpdir/public.jsp" target="_blank">Federal depository library</a>.</div><hr class="black-1-cccccc" /><div>    <a href="https://www.govinfo.gov"><img alt="Discover GPO's next generation of government information! Start searching
and browsing"src="/fdsysgpopages/images/govinfo-release-banner-200x200.png"></div></div></td><td class="vert-spacer-450">&nbsp;</td><td class="two-col-layout-right"><span id="breadcrumbs"><a href="search/home.action">FDsys</a> &gt;<span class="nowrap"><a
href="/fdsys/browse/collectiontab.action?collectionCode=CHRG&amp;bread=true">Browse By Collection</a> &gt; </span><span class="nowrap"><a href="/fdsys/browse/collection.action?collectionCode=CHRG&amp;bread=true">Collection Results</a> </span>  </span><!-- end of 'breadcrumbs' --><div
id="browse-layout-mask"><div>&nbsp;</div><h3 class="page-title">Congressional Hearings</h3><p>  A hearing is a meeting or session of a Senate, House, joint, or special committee of Congress,usually open to the public, to obtain information and opinions on proposed legislation,conduct an
investigation, or evaluate/oversee the activities of a government departmentor the implementation of a Federal law. In addition, hearings may also be purely exploratoryin nature, providing testimony and data about topics of current interest.Most congressional hearings are published two months
to two years after they are held.</p><p>Not all congressional hearings are available on FDsys. Whether or not a hearing is disseminated on FDsys depends on the committee. GPO continues to add hearings as they become available during each session of Congress. If a congressional hearing is not
listed in FDsys, it is not available electronically via GPO at this time. NOTE: If a committee has not made a hearing available electronically via GPO for a specific Congress, the committee's name will not appear in the browse list until a hearing for that committee is made available in
FDsys. Committee publications may also be browsed via the FDsys <a href="/fdsys/browse/committeetab.action">congressional committee browse page</a>.</p><p> FDsys contains selected House and Senate hearings for the 105th Congress(1997-98) forward. The House and Senate appropriations hearings
for fiscal year 1998 forward are also included. Documents are available as text and PDF. Graphics, including scanned images of camera ready copy are omitted from the ASCII text version but are contained in the PDF files.<a href="javascript:popUpHelp(BASE_USER_HELP + helpLinks.CHRG);">About
the Congressional Hearings</a>.</p><div id="browse-drilldown-mask"><input type="hidden" name="urlhid" value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=115&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord"
value="" id="ycord"/><div class="level1 browse-level"><a href="#" onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=115&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />115th Congress (2017 - 2018)
</a></div><input type="hidden" name="urlhid" value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=114&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a
href="#" onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=114&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />114th Congress (2015 - 2016)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=113&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=113&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />113th Congress (2013 - 2014)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=112&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=112&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />112th Congress (2011 - 2012)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=111&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=111&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />111th Congress (2009 - 2010)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=110&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=110&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />110th Congress (2007 - 2008)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=109&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=109&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />109th Congress (2005 - 2006)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=108&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=108&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />108th Congress (2003 - 2004)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=107&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=107&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />107th Congress (2001 - 2002)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=106&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=106&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />106th Congress (1999 - 2000)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=105&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=105&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />105th Congress (1997 - 1998)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=104&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=104&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />104th Congress (1995 - 1996)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=103&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=103&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />103rd Congress (1993 - 1994)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=102&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=102&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />102nd Congress (1991 - 1992)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=100&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=100&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />100th Congress (1987 - 1988)</a></div><input type="hidden" name="urlhid"
value="/fdsys/browse/collection.action?collectionCode=CHRG&amp;amp;browsePath=99&amp;amp;isCollapsed=false&amp;amp;leafLevelBrowse=false" id="urlhid1"/><input type="hidden" name="ycord" value="" id="ycord"/><div class="level1 browse-level"><a href="#"
onclick="goWithVars('/fdsys/browse/collection.action?collectionCode=CHRG&amp;browsePath=99&amp;isCollapsed=false&amp;leafLevelBrowse=false',''); return false;"><img alt="+" src="images/plus.gif" />99th Congress (1985 - 1986)</a></div></div></div><!-- End: browse-layout-mask --></td></tr>
</table><script type="text/javascript">var addthis_config = {"data_track_clickback":true, "ui_508_compliant":true};</script><script type="text/javascript" src="//s7.addthis.com/js/300/addthis_widget.js#pubid=ra-51ad068b342db0bf"></script><div id="footer" style="background-color:white;"> <div
id="gpo-address-bar" style="background-color:white;">732 North Capitol Street, NW, Washington, DC 20401-0001&nbsp;&nbsp;202.512.1800&nbsp;&nbsp;    </div>  <div id="footer-links" style="background-color:white;"><a href="https://github.com/usgpo">GitHub</a>&nbsp;<span class="footer-pipe">|
</span>&nbsp;<a href="http://www.gpo.gov/smap/fdsys/sitemap.xml">Sitemap</a>&nbsp;<span class="footer-pipe">|</span>&nbsp;<a href="http://www.gpo.gov/fdsys/bulkdata">Bulk Data</a>&nbsp;<span class="footer-pipe">|</span>&nbsp;<a href="http://www.gpo.gov/etc/privacy.htm">Privacy</a>&nbsp;<span
class="footer-pipe">|</span>&nbsp;<a href="http://www.gpo.gov/etc/section-508.htm">Accessibility</a></div>  <br/>   <div class="addthis_toolbox addthis_default_style" style="background-color:white; padding-left:10px;">  <a class="addthis_button_preferred_1"></a>  <a
class="addthis_button_preferred_2"></a> <a class="addthis_button_preferred_3"></a>  <a class="addthis_button_preferred_4"></a>  <a class="addthis_button_compact"></a>  </div><br/></div>   <!-- Start SDC Tag -->  <!-- The scripts below provide functionality for web trends stastics--> <script
type="text/javascript" src="https://www.gpo.gov/fdsys/scripts/wtinit.js"></script>  <script type="text/javascript" src="https://www.gpo.gov/fdsys/scripts/wtbase.js"></script><!-- NoScript tag below provides a way to capture web trends information for those browsers that  dont hava java script
enabled in their browsers--><noscript><div style="display: none;"><img alt="DCSIMG" id="DCSIMG" width="1" height="1"src="http://analytics.gpo.gov/dcsjsw8h600000gotf0vyrmly_2j7v/njs.gif?dcsuri=/nojavascript&amp;WT.js=No&amp;dcscfg=1&amp;WT.tv=1.3.7" /></div>   </noscript><!--End SDC Tag -->
</div><!-- wrapper --></body></html>
```

HTML PRIMER
TAGS + ATTRIBUTES

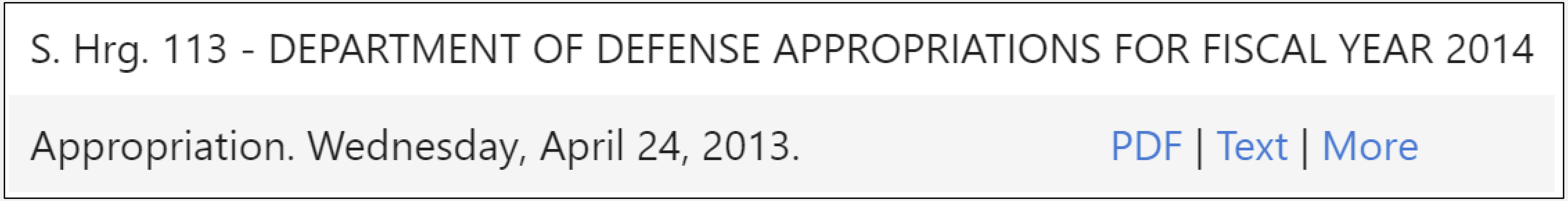


Tag name	Function
`<head>`	The head of a page includes general information such as its authors, language etc.
`<body>`	All the content you see in the browser is inside the body tags.
`<p>`	A paragraph of text.
`<br>`	A break between two lines of text.
`<b>` , `<i>` , `<u>`	Display text in bold, italics or underlined.
`<a href="gpo.html">`	A link to the page gpo.html.
`<img src="gpo.jpg">`	Displays the picture gpo.jpg.
`<table>` , `<tr>` , `<td>`	A table including its rows (tr) and cells (td).
`<div>`	Divides an HTML document into different parts. Usually used for layout purposes.
`<span>`	Similar to `<div>` tag, but only for small parts of the code.

attribute

HTML PRIMER

DOCUMENT STRUCTURE



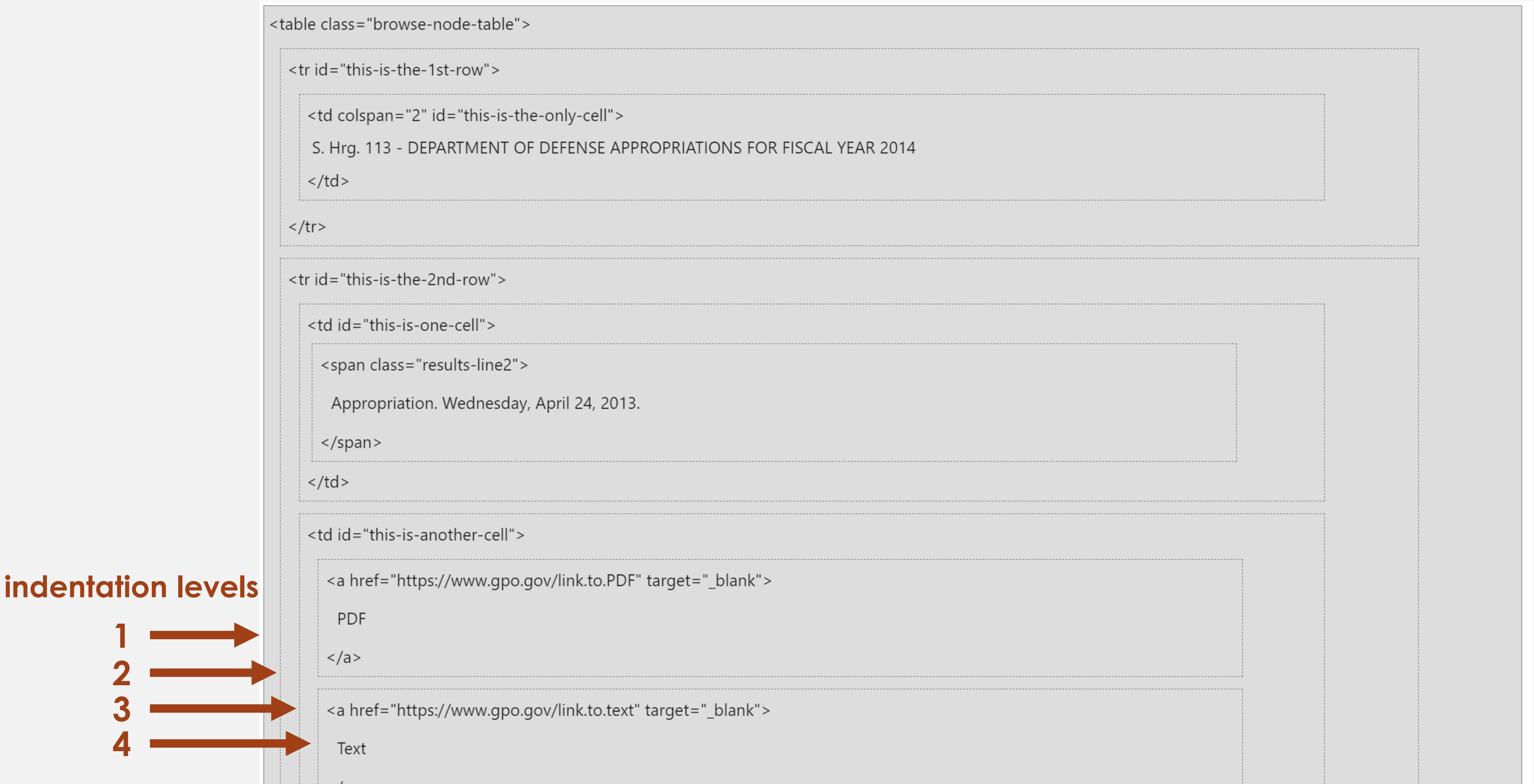
S. Hrg. 113 - DEPARTMENT OF DEFENSE APPROPRIATIONS FOR FISCAL YEAR 2014

Appropriation. Wednesday, April 24, 2013. PDF | Text | More

```
<table class="browse-node-table">
<tr id="this-is-the-1st-row">
<td colspan="2" id="this-is-the-only-cell"> S. Hrg. 113 - DEPARTMENT OF DEFENSE
APPROPRIATIONS FOR FISCAL YEAR 2014 </td> </tr>
<tr id="this-is-the-2nd-row">
<td id="this-is-one-cell"> <span class="results-line2"> Appropriation. Wednesday, April 24,
2013. </span> </td>
 <td id="this-is-another-cell"> <a href="https://www.gpo.gov/link.to.PDF" target="_blank">
PDF </a> <a href="https://www.gpo.gov/link.to.text" target="_blank"> Text </a> <a
href="https://www.gpo.gov/link.to.more" target="_blank"> More </a> </td> </tr> <tr
id="this-is-the-3rd-row"> <td colspan="2" id="this-is-the-only-cell"> </td> </tr>
 </table>
```

HTML PRIMER

DOCUMENT STRUCTURE - INDENTED



```
<table class="browse-node-table">
  <tr id="this-is-the-1st-row">
    <td colspan="2" id="this-is-the-only-cell">
     S. Hrg. 113 - DEPARTMENT OF DEFENSE APPROPRIATIONS FOR FISCAL YEAR 2014
    </td>
  </tr>
  <tr id="this-is-the-2nd-row">
    <td id="this-is-one-cell">
      <span class="results-line2">
        Appropriation. Wednesday, April 24, 2013.
      </span>
    </td>
    <td id="this-is-another-cell">
      <a href="https://www.gpo.gov/link.to.PDF" target="_blank">
        PDF
      </a>
      <a href="https://www.gpo.gov/link.to.text" target="_blank">
        Text
```

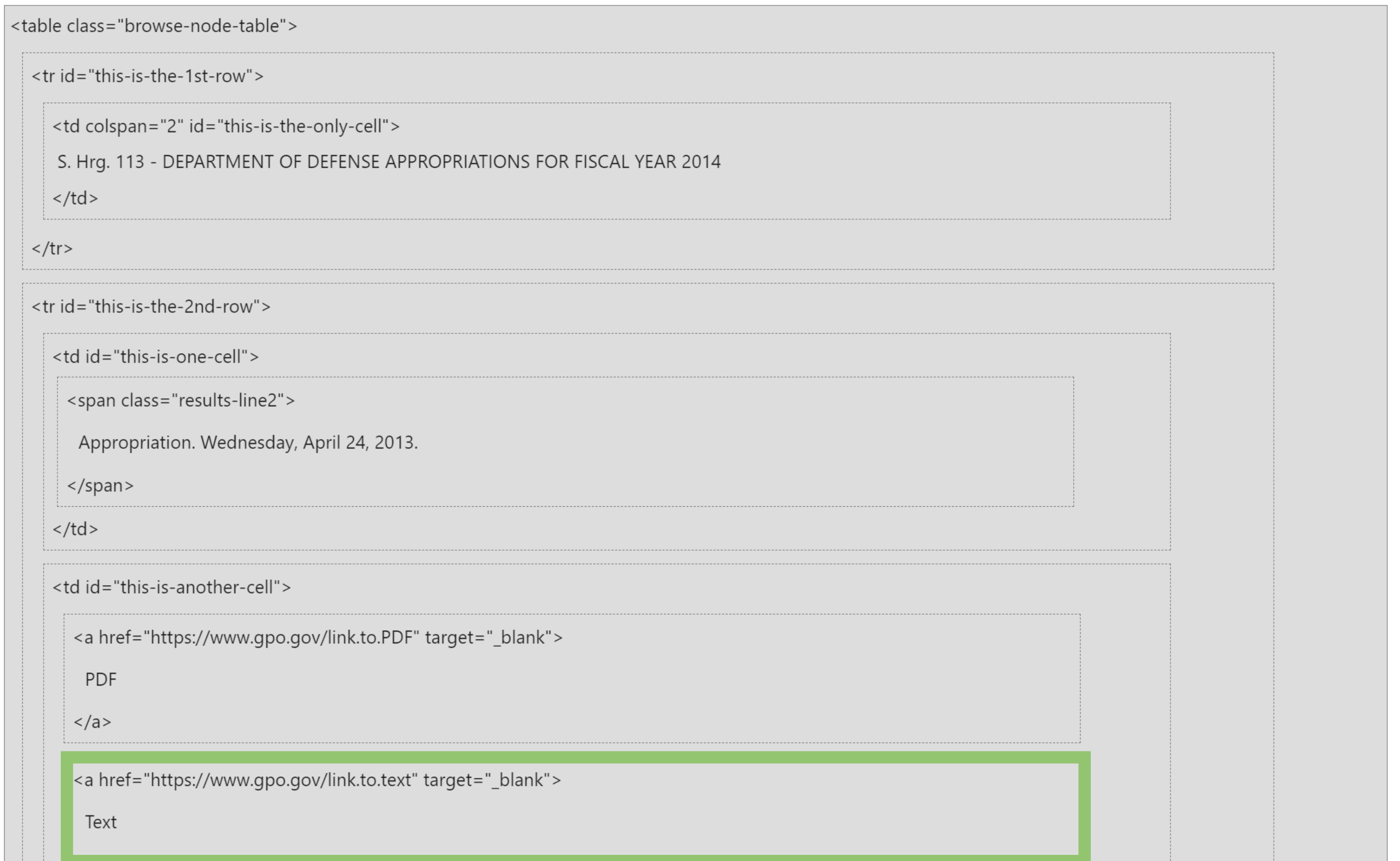
XPath
Navigating the DOM

- ► Simple query language to describe DOM locations.

- ► Basic rules:
  1. We only include the opening tags. The program will just ignore closing tags it sees inside the code along the way.
  2. The path only includes tags where the program has to move into a new nest (or towards the right in the above visualization).
  3. We use numbers in squared brackets to identify a specific tag in cases where there are multiple ones of the same kind within the same nest.
  4. We connect the different pieces of our path with backslashes in-between each tag.

XPATH EXAMPLE

CLICKING ON THE «TEXT» LINK



```
<table class="browse-node-table">
  <tr id="this-is-the-1st-row">
    <td colspan="2" id="this-is-the-only-cell">
     S. Hrg. 113 - DEPARTMENT OF DEFENSE APPROPRIATIONS FOR FISCAL YEAR 2014
    </td>
  </tr>
  <tr id="this-is-the-2nd-row">
    <td id="this-is-one-cell">
      <span class="results-line2">
        Appropriation. Wednesday, April 24, 2013.
      </span>
    </td>
    <td id="this-is-another-cell">
      <a href="https://www.gpo.gov/link.to.PDF" target="_blank">
        PDF
      </a>
      <a href="https://www.gpo.gov/link.to.text" target="_blank">
        Text
```

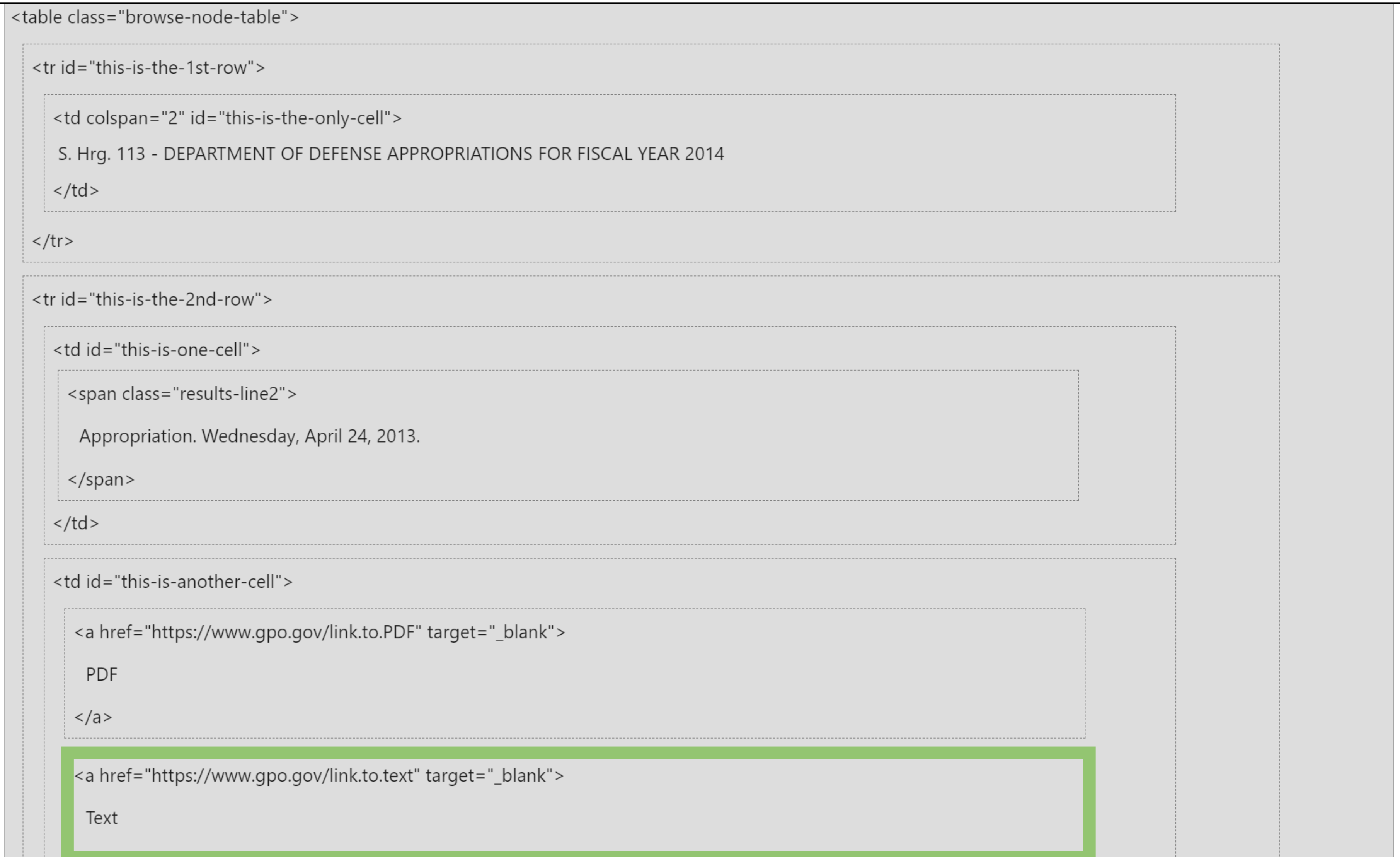
``table`` "Start at the beginning of our code (``table``),

``tr[2]`` use the second row tag (``tr[2]``).

``td[2]`` Within that row tag, go into the second cell tag you can find (``td[2]``), and

``a[2]``stop in that cell's second link tag (``a[2]``)."

➔ ``/table/tr[2]/td[2]/a[2]``



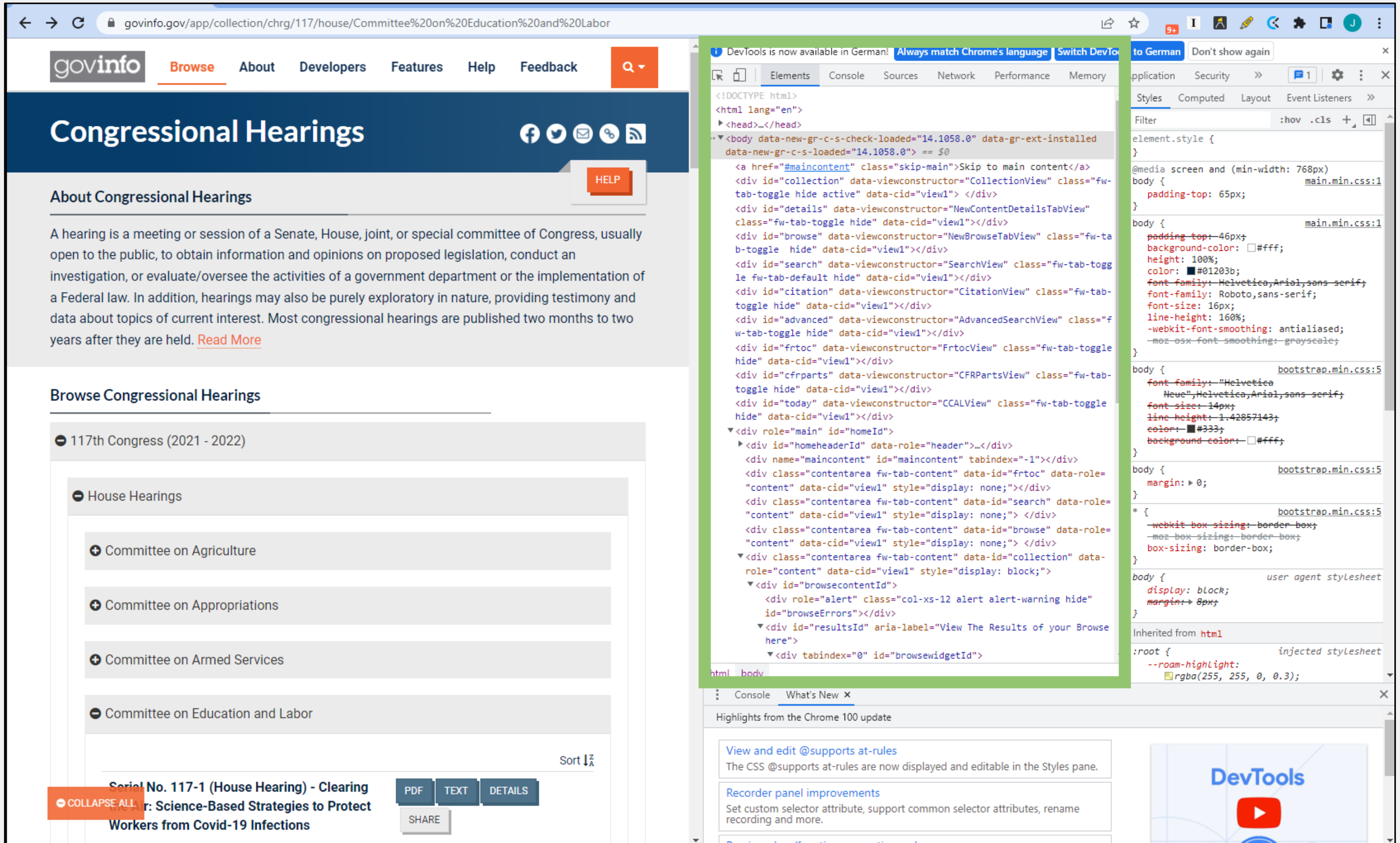
```
<table class="browse-node-table">
  <tr id="this-is-the-1st-row">
    <td colspan="2" id="this-is-the-only-cell">
     S. Hrg. 113 - DEPARTMENT OF DEFENSE APPROPRIATIONS FOR FISCAL YEAR 2014
    </td>
  </tr>
  <tr id="this-is-the-2nd-row">
    <td id="this-is-one-cell">
      <span class="results-line2">
        Appropriation. Wednesday, April 24, 2013.
      </span>
    </td>
    <td id="this-is-another-cell">
      <a href="https://www.gpo.gov/link.to.PDF" target="_blank">
        PDF
      </a>
      <a href="https://www.gpo.gov/link.to.text" target="_blank">
        Text
```

FINDING THE XPATH

USING THE «INSPECTOR» IN CHROME/FIREFOX

Open with:

- ▶ CTRL + SHIFT + I
- ▶ ⌥ + ⌘ + I



govinfo.gov/app/collection/chrg/117/house/Committee%20on%20Education%20and%20Labor

Congressional Hearings

About Congressional Hearings

A hearing is a meeting or session of a Senate, House, joint, or special committee of Congress, usually open to the public, to obtain information and opinions on proposed legislation, conduct an investigation, or evaluate/oversee the activities of a government department or the implementation of a Federal law. In addition, hearings may also be purely exploratory in nature, providing testimony and data about topics of current interest. Most congressional hearings are published two months to two years after they are held. Read More

Browse Congressional Hearings

- ⊖ 117th Congress (2021 - 2022)
  - ⊖ House Hearings
    - ⊕ Committee on Agriculture
    - ⊕ Committee on Appropriations
    - ⊕ Committee on Armed Services
    - ⊖ Committee on Education and Labor

Sort

Serial No. 117-1 (House Hearing) - Clearing the Air: Science-Based Strategies to Protect Workers from Covid-19 Infections

DevTools is now available in German! Always match Chrome's language Switch DevTools to German Don't show again

```
<!DOCTYPE html>
<html lang="en">
 ▶<head>…</head>
··▼<body data-new-gr-c-s-check-loaded="14.1058.0" data-gr-ext-installed
    data-new-gr-c-s-loaded="14.1058.0"> == $0
      <a href="#maincontent" class="skip-main">Skip to main content</a>
      <div id="collection" data-viewconstructor="CollectionView" class="fw-
      tab-toggle hide active" data-cid="view1"> </div>
      <div id="details" data-viewconstructor="NewContentDetailsTabView"
      class="fw-tab-toggle hide" data-cid="view1"></div>
      <div id="browse" data-viewconstructor="NewBrowseTabView" class="fw-ta
      b-toggle  hide" data-cid="view1"></div>
      <div id="search" data-viewconstructor="SearchView" class="fw-tab-togg
      le fw-tab-default hide" data-cid="view1"></div>
      <div id="citation" data-viewconstructor="CitationView" class="fw-tab-
      toggle hide" data-cid="view1"></div>
      <div id="advanced" data-viewconstructor="AdvancedSearchView" class="f
      w-tab-toggle hide" data-cid="view1"></div>
      <div id="frtoc" data-viewconstructor="FrtocView" class="fw-tab-toggle
      hide" data-cid="view1"></div>
      <div id="cfrparts" data-viewconstructor="CFRPartsView" class="fw-tab-
      toggle hide" data-cid="view1"></div>
      <div id="today" data-viewconstructor="CCALView" class="fw-tab-toggle
      hide" data-cid="view1"></div>
    ▼<div role="main" id="homeId">
      ▶<div id="homeheaderId" data-role="header">…</div>
       <div name="maincontent" id="maincontent" tabindex="-1"></div>
       <div class="contentarea fw-tab-content" data-id="frtoc" data-role=
       "content" data-cid="view1" style="display: none;"></div>
       <div class="contentarea fw-tab-content" data-id="search" data-role=
       "content" data-cid="view1" style="display: none;"> </div>
       <div class="contentarea fw-tab-content" data-id="browse" data-role=
       "content" data-cid="view1" style="display: none;"> </div>
      ▼<div class="contentarea fw-tab-content" data-id="collection" data-
       role="content" data-cid="view1" style="display: block;">
        ▼<div id="browsecontentId">
           <div role="alert" class="col-xs-12 alert alert-warning hide"
           id="browseErrors"></div>
          ▼<div id="resultsId" aria-label="View The Results of your Browse
           here">
            ▼<div tabindex="0" id="browsewidgetId">
html  body
```

```
element.style {
}
@media screen and (min-width: 768px)
body {                                  main.min.css:1
   padding-top: 65px;
}
body {                                  main.min.css:1
   ~~padding-top: 46px;~~
   background-color: □#fff;
   height: 100%;
   color: ■#01203b;
   ~~font-family: Helvetica,Arial,sans-serif;~~
   font-family: Roboto,sans-serif;
   font-size: 16px;
   line-height: 160%;
   -webkit-font-smoothing: antialiased;
   ~~-moz-osx-font-smoothing: grayscale;~~
}
body {                             bootstrap.min.css:5
   ~~font-family: "Helvetica~~
      ~~Neue",Helvetica,Arial,sans-serif;~~
   ~~font-size: 14px;~~
   ~~line-height: 1.42857143;~~
   ~~color: ■#333;~~
   ~~background-color: □#fff;~~
}
body {                             bootstrap.min.css:5
   margin: ▶ 0;
}
* {                                bootstrap.min.css:5
   ~~-webkit-box-sizing: border-box;~~
   ~~-moz-box-sizing: border-box;~~
   box-sizing: border-box;
}
body {                          user agent stylesheet
   display: block;
   ~~margin: ▶ 8px;~~
}
Inherited from html
:root {                           injected stylesheet
   --roam-highlight:
      □rgba(255, 255, 0, 0.3);
```

Highlights from the Chrome 100 update

View and edit @supports at-rules

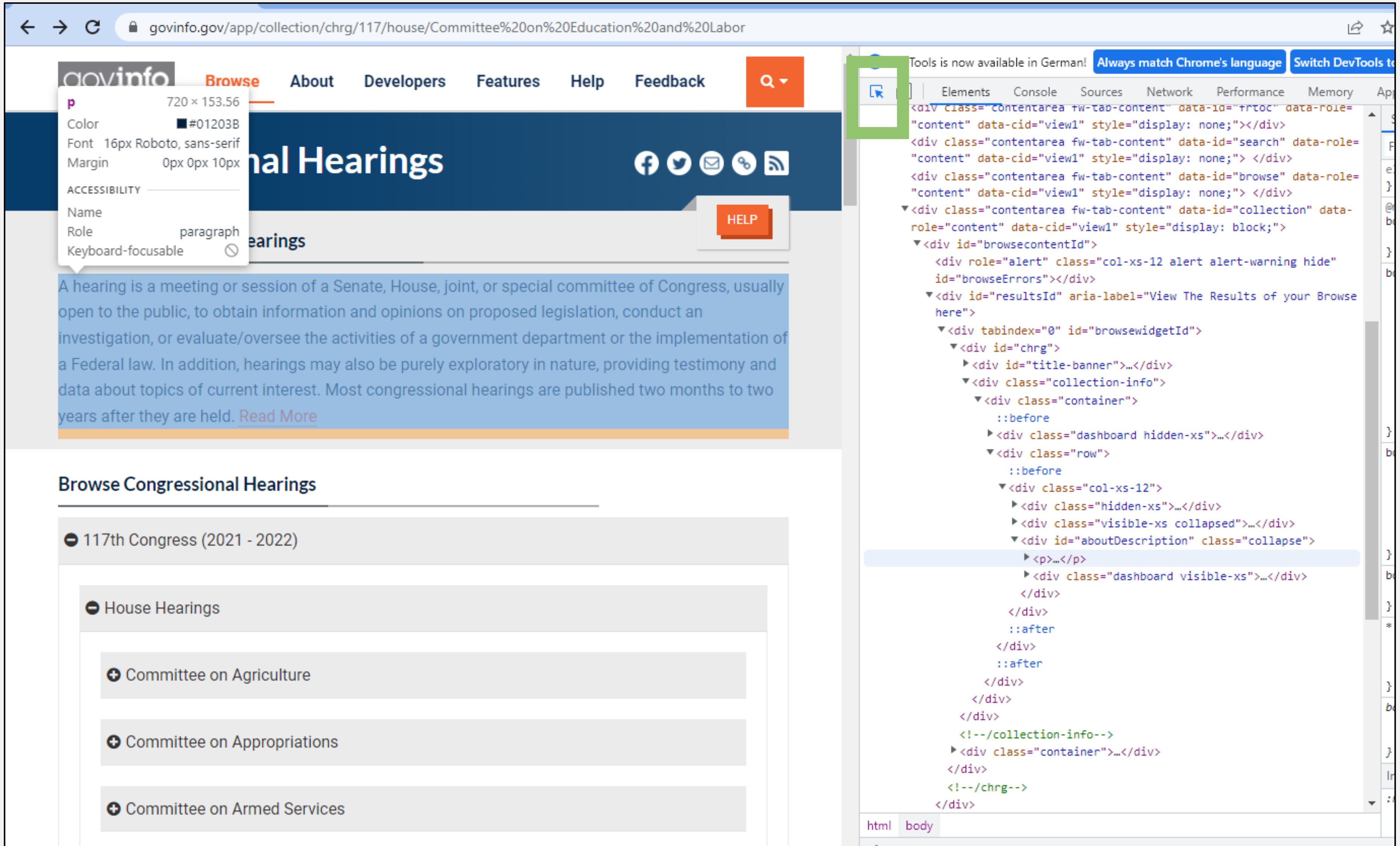
The CSS @supports at-rules are now displayed and editable in the Styles pane.

Recorder panel improvements

Set custom selector attribute, support common selector attributes, rename recording and more.

Finding the XPath

Select element



govinfo.gov/app/collection/chrg/117/house/Committee%20on%20Education%20and%20Labor
govinfo
Browse
About
Developers
Features
Help
Feedback
nal Hearings
HELP
earings
Tools is now available in German!
Always match Chrome's language
Switch DevTools t
Elements
Console
Sources
Network
Performance
Memory
Ap
html
body

p 720 × 153.56

Color ■#01203B

Font 16px Roboto, sans-serif

Margin 0px 0px 10px

ACCESSIBILITY

Name

Role paragraph

Keyboard-focusable ⃠

A hearing is a meeting or session of a Senate, House, joint, or special committee of Congress, usually open to the public, to obtain information and opinions on proposed legislation, conduct an investigation, or evaluate/oversee the activities of a government department or the implementation of a Federal law. In addition, hearings may also be purely exploratory in nature, providing testimony and data about topics of current interest. Most congressional hearings are published two months to two years after they are held. Read More

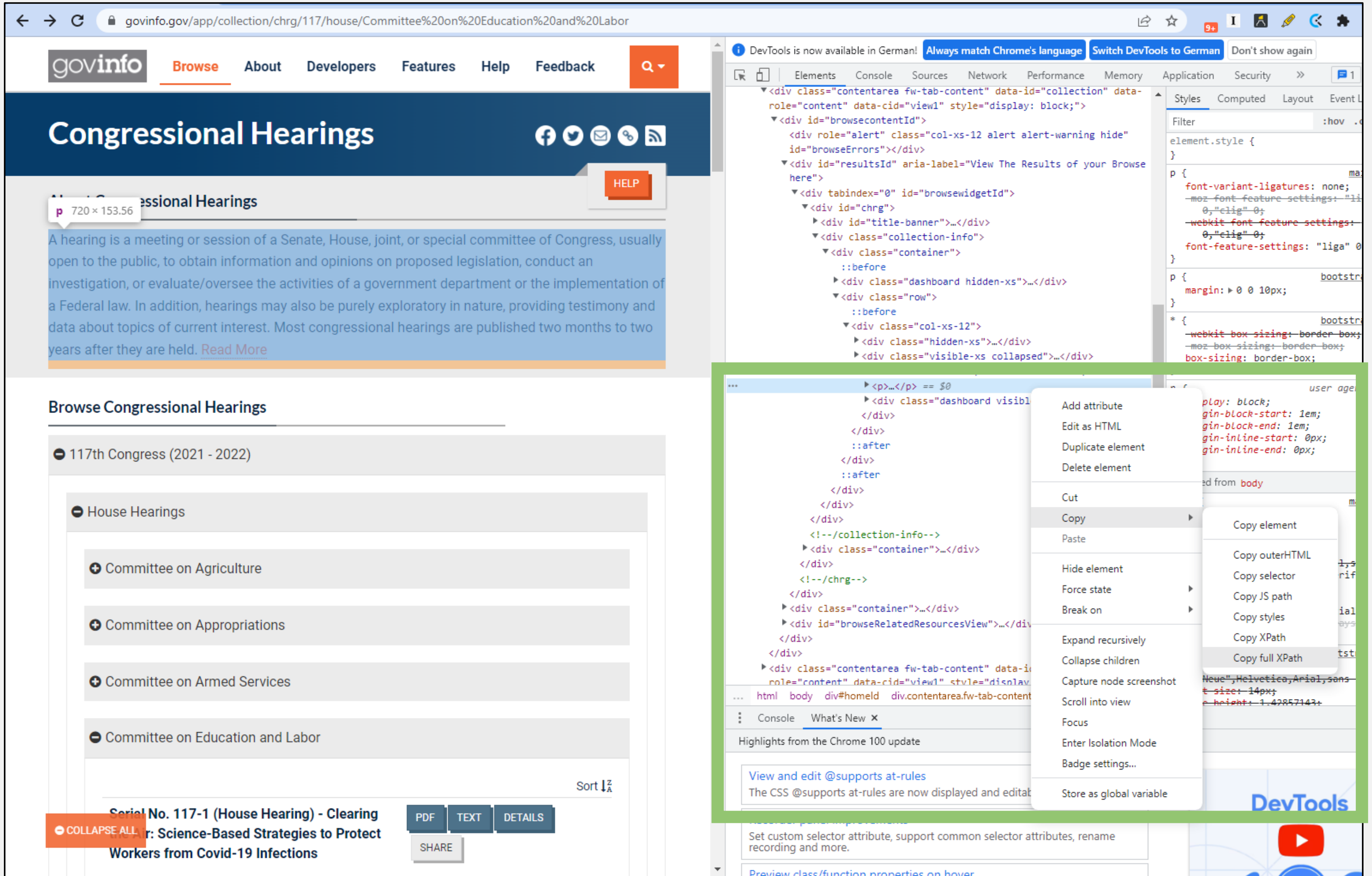
Browse Congressional Hearings

- 117th Congress (2021 - 2022)
  - House Hearings
    - Committee on Agriculture
    - Committee on Appropriations
    - Committee on Armed Services

```
<div class="contentarea fw-tab-content" data-id="frtoc" data-role=
"content" data-cid="view1" style="display: none;"></div>
<div class="contentarea fw-tab-content" data-id="search" data-role=
"content" data-cid="view1" style="display: none;"> </div>
<div class="contentarea fw-tab-content" data-id="browse" data-role=
"content" data-cid="view1" style="display: none;"> </div>
▼<div class="contentarea fw-tab-content" data-id="collection" data-
role="content" data-cid="view1" style="display: block;">
  ▼<div id="browsecontentId">
      <div role="alert" class="col-xs-12 alert alert-warning hide"
      id="browseErrors"></div>
    ▼<div id="resultsId" aria-label="View The Results of your Browse
      here">
      ▼<div tabindex="0" id="browsewidgetId">
        ▼<div id="chrg">
          ▶<div id="title-banner">…</div>
          ▼<div class="collection-info">
            ▼<div class="container">
                ::before
              ▶<div class="dashboard hidden-xs">…</div>
              ▼<div class="row">
                  ::before
                ▼<div class="col-xs-12">
                  ▶<div class="hidden-xs">…</div>
                  ▶<div class="visible-xs collapsed">…</div>
                  ▼<div id="aboutDescription" class="collapse">
                    ▶<p>…</p>
                    ▶<div class="dashboard visible-xs">…</div>
                    </div>
                  </div>
                  ::after
                </div>
                ::after
              </div>
            </div>
          </div>
          <!--/collection-info-->
        ▶<div class="container">…</div>
        </div>
        <!--/chrg-->
      </div>
```

Finding the XPath

Copy preferred XPath version



govinfo.gov/app/collection/chrg/117/house/Committee%20on%20Education%20and%20Labor
govinfo
Browse
About
Developers
Features
Help
Feedback
Congressional Hearings
HELP
p 720 × 153.56
About Congressional Hearings

A hearing is a meeting or session of a Senate, House, joint, or special committee of Congress, usually open to the public, to obtain information and opinions on proposed legislation, conduct an investigation, or evaluate/oversee the activities of a government department or the implementation of a Federal law. In addition, hearings may also be purely exploratory in nature, providing testimony and data about topics of current interest. Most congressional hearings are published two months to two years after they are held. Read More

Browse Congressional Hearings
117th Congress (2021 - 2022)
House Hearings
Committee on Agriculture
Committee on Appropriations
Committee on Armed Services
Committee on Education and Labor
Sort
PDF
TEXT
DETAILS
SHARE
COLLAPSE ALL

Serial No. 117-1 (House Hearing) - Clearing the Air: Science-Based Strategies to Protect Workers from Covid-19 Infections

DevTools is now available in German!

```
▼<div class="contentarea fw-tab-content" data-id="collection" data-
role="content" data-cid="view1" style="display: block;">
 ▼<div id="browsecontentId">
   <div role="alert" class="col-xs-12 alert alert-warning hide"
   id="browseErrors"></div>
  ▼<div id="resultsId" aria-label="View The Results of your Browse
   here">
   ▼<div tabindex="0" id="browsewidgetId">
    ▼<div id="chrg">
     ▶<div id="title-banner">…</div>
     ▼<div class="collection-info">
      ▼<div class="container">
         ::before
       ▶<div class="dashboard hidden-xs">…</div>
       ▼<div class="row">
          ::before
        ▼<div class="col-xs-12">
         ▶<div class="hidden-xs">…</div>
         ▶<div class="visible-xs collapsed">…</div>
...      ▶<p>…</p> == $0
         ▶<div class="dashboard visibl
          </div>
        </div>
        ::after
       </div>
       ::after
      </div>
     </div>
    </div>
    <!--/collection-info-->
   ▶<div class="container">…</div>
   </div>
   <!--/chrg-->
  </div>
 ▶<div class="container">…</div>
 ▶<div id="browseRelatedResourcesView">…</div
 </div>
</div>
▶<div class="contentarea fw-tab-content" data-i
role="content" data-cid="view1" style="display
```

```
element.style {
}
p {
  font-variant-ligatures: none;
  ~~-moz-font-feature-settings: "li~~
      ~~0,"clig" 0;~~
  ~~-webkit-font-feature-settings:~~
      ~~0,"clig" 0;~~
  font-feature-settings: "liga" 0
}
p {                          bootstr
  margin: ▶ 0 0 10px;
}
* {                          bootstr
  ~~-webkit-box-sizing: border-box;~~
  ~~-moz-box-sizing: border-box;~~
  box-sizing: border-box;
}
                             user age
play: block;
gin-block-start: 1em;
gin-block-end: 1em;
gin-inline-start: 0px;
gin-inline-end: 0px;
ed from body
Neue",Helvetica,Arial,sans
t-size: 14px;
-height: 1.42857143;
```

Highlights from the Chrome 100 update

View and edit @supports at-rules
The CSS @supports at-rules are now displayed and editab

Set custom selector attribute, support common selector attributes, rename recording and more.

Preview class/function properties on hover

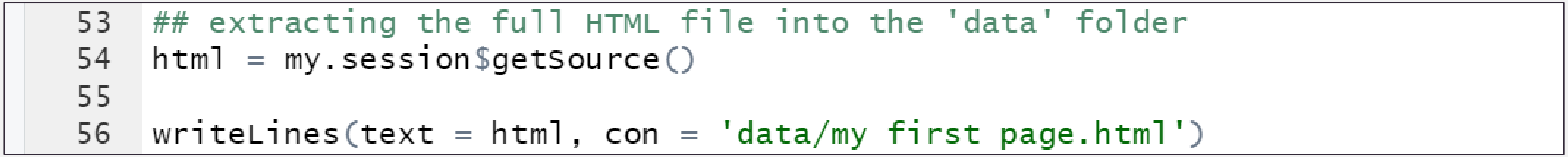
Short XPaths protect against HTML coding errors

- Full XPath
  /html/body/div[10]/div[6]/div/div[2]/div/div[1]/div[2]/div/div[2]/div/div[3]/p

- Short XPath
  //*[@id="aboutDescription"]/p

- CSS selector (less flexible)
  #aboutDescription > p

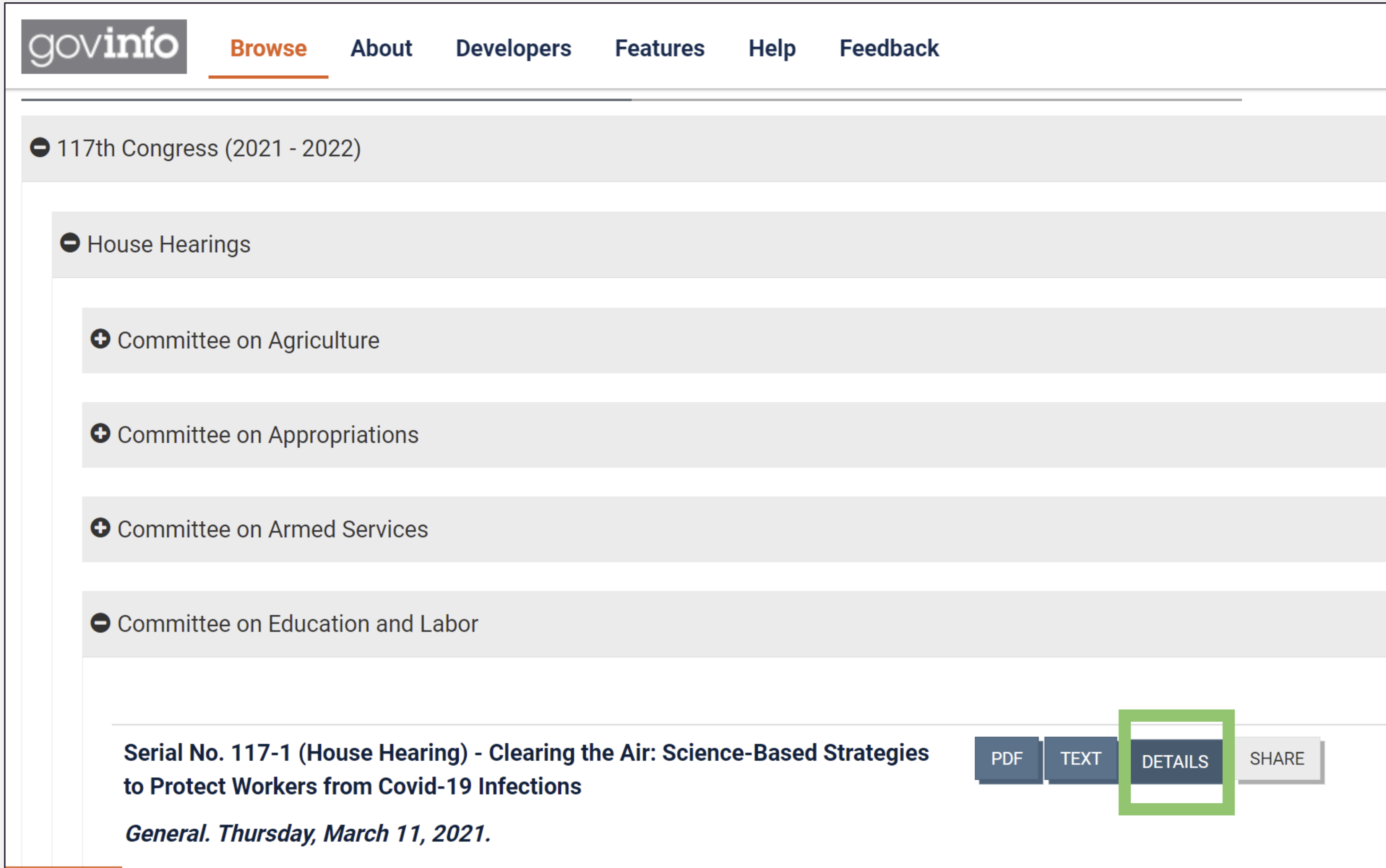
Adding to our script: navigation + storage



```
## extracting the full HTML file into the 'data' folder
html = my.session$getSource()

writeLines(text = html, con = 'data/my first page.html')
```

2. Create a local copy

Can you capture this details page?



117th Congress (2021 - 2022)

House Hearings

Committee on Agriculture

Committee on Appropriations

Committee on Armed Services

Committee on Education and Labor

Serial No. 117-1 (House Hearing) - Clearing the Air: Science-Based Strategies to Protect Workers from Covid-19 Infections

General. Thursday, March 11, 2021.

check your result: url

HARVESTING ONLINE DATA

Short course at Santa Clara University, ECON 173

Johannes Fritz

5/12/22

YOU WILL LEARN 3 TOOLS

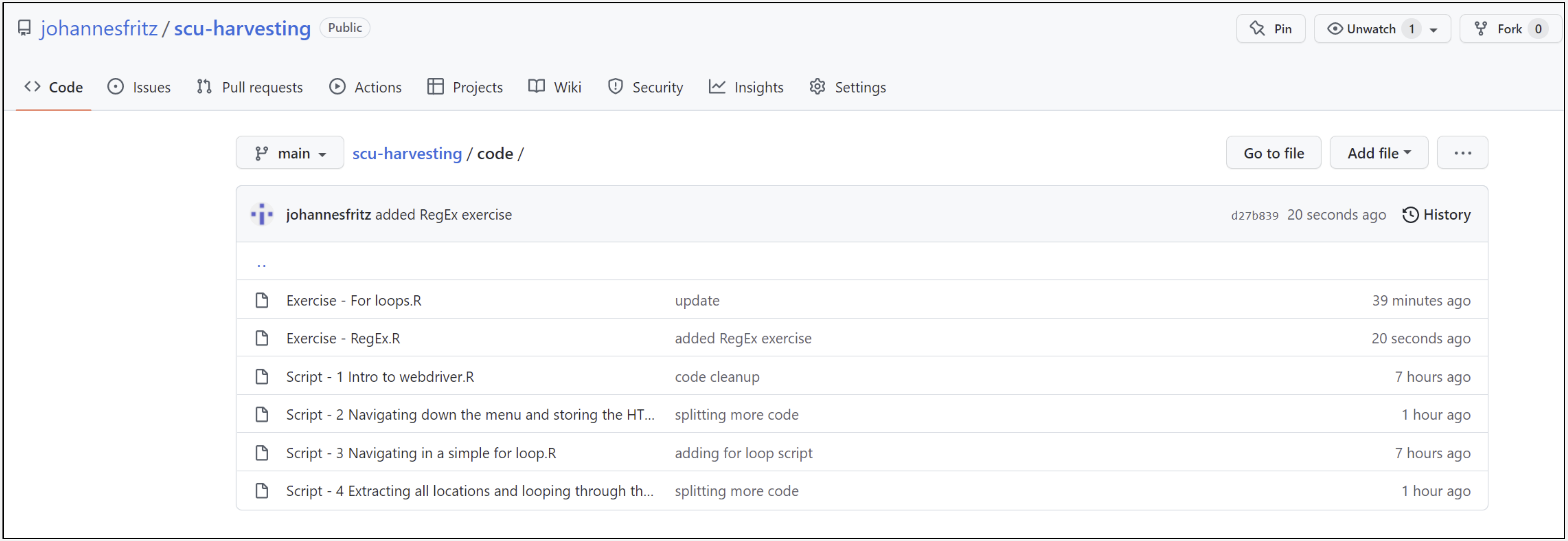
1. Harvesting the data:
   Headless **web scraping** using R.

2. Parsing the data:
   Introduction to **Regular Expressions**.

3. Organizing the data:
   Basics of **relational database design**.

All material at:
https://github.com/johannesfritz/scu-harvesting

RE-ORGANISED THE CODE INTO THE STEPS WE TOOK

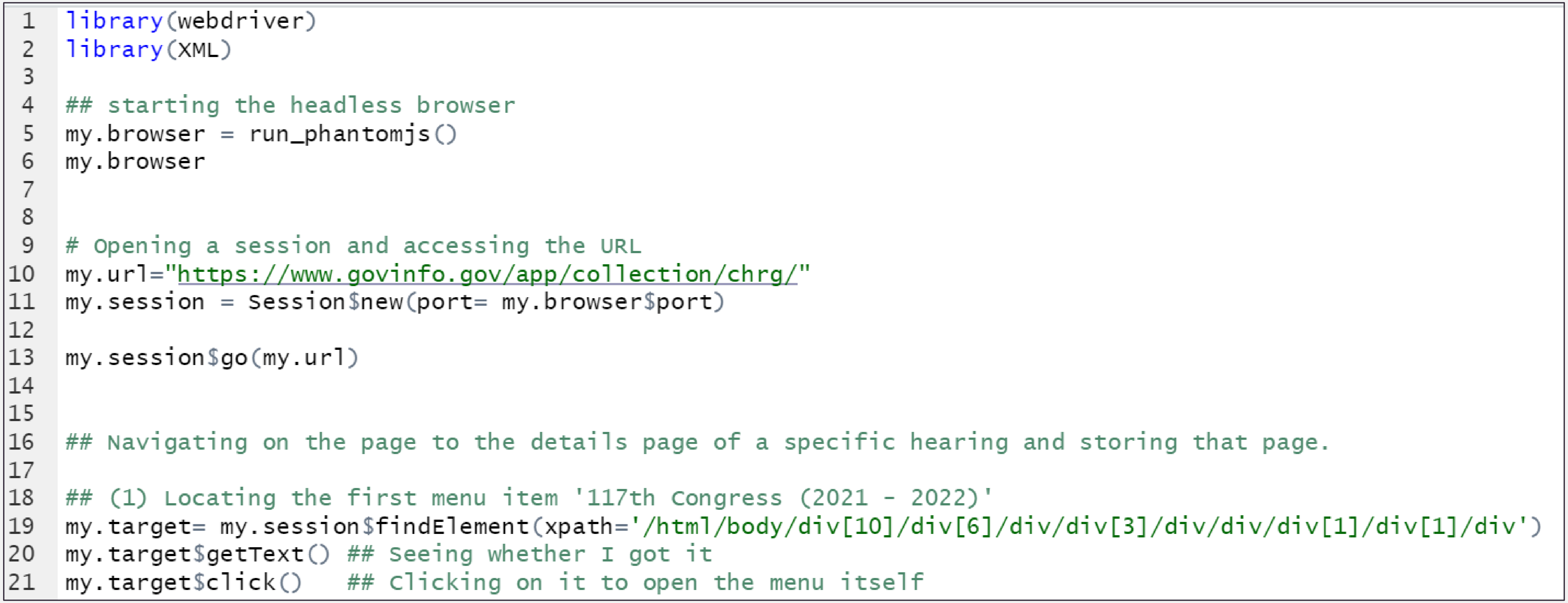


johannesfritz added RegEx exercise d27b839 20 seconds ago

..

Exercise - For loops.R	update	39 minutes ago
Exercise - RegEx.R	added RegEx exercise	20 seconds ago
Script - 1 Intro to webdriver.R	code cleanup	7 hours ago
Script - 2 Navigating down the menu and storing the HT...	splitting more code	1 hour ago
Script - 3 Navigating in a simple for loop.R	adding for loop script	7 hours ago
Script - 4 Extracting all locations and looping through th...	splitting more code	1 hour ago

FROM LAST WEEK



```r
library(webdriver)
library(XML)

## starting the headless browser
my.browser = run_phantomjs()
my.browser


# Opening a session and accessing the URL
my.url="https://www.govinfo.gov/app/collection/chrg/"
my.session = Session$new(port= my.browser$port)

my.session$go(my.url)


## Navigating on the page to the details page of a specific hearing and storing that page.

## (1) Locating the first menu item '117th Congress (2021 - 2022)'
my.target= my.session$findElement(xpath='/html/body/div[10]/div[6]/div/div[3]/div/div/div[1]/div[1]/div')
my.target$getText() ## Seeing whether I got it
my.target$click()   ## Clicking on it to open the menu itself
```

More XPath: axes

- ► If you have a prominent position in the code (e.g. an attribute).
- ► Axes allow you to navigate from there more flexibly without stating the path
- ► Axes names
  - > parent or child
  - > ancestor or descendant
  - > preceding-sibling or following-sibling

HTML PRIMER

DOCUMENT STRUCTURE - INDENTED

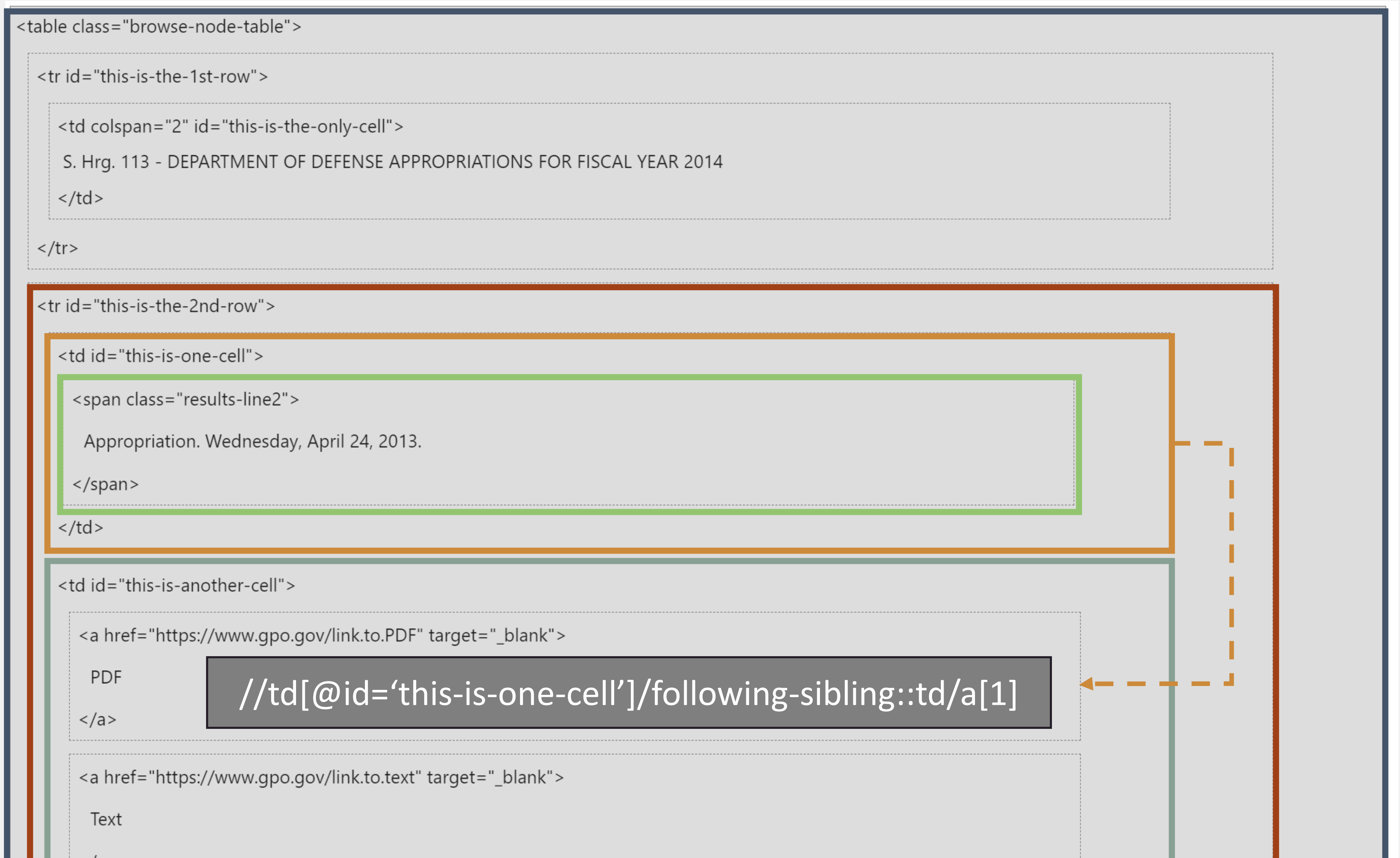
ancestor

parent

start

child

following-sibling



```
<table class="browse-node-table">
  <tr id="this-is-the-1st-row">
    <td colspan="2" id="this-is-the-only-cell">
     S. Hrg. 113 - DEPARTMENT OF DEFENSE APPROPRIATIONS FOR FISCAL YEAR 2014
    </td>
  </tr>
  <tr id="this-is-the-2nd-row">
    <td id="this-is-one-cell">
      <span class="results-line2">
        Appropriation. Wednesday, April 24, 2013.
      </span>
    </td>
    <td id="this-is-another-cell">
      <a href="https://www.gpo.gov/link.to.PDF" target="_blank">
        PDF
      </a>
      <a href="https://www.gpo.gov/link.to.text" target="_blank">
        Text
```

//td[@id=‘this-is-one-cell’]/following-sibling::td/a[1]

XPATH RESOURCE

W3schools: XPath tutorial
https://www.w3schools.com/xml/xpath_intro.asp

- Syntax
- Axes
- Special operators

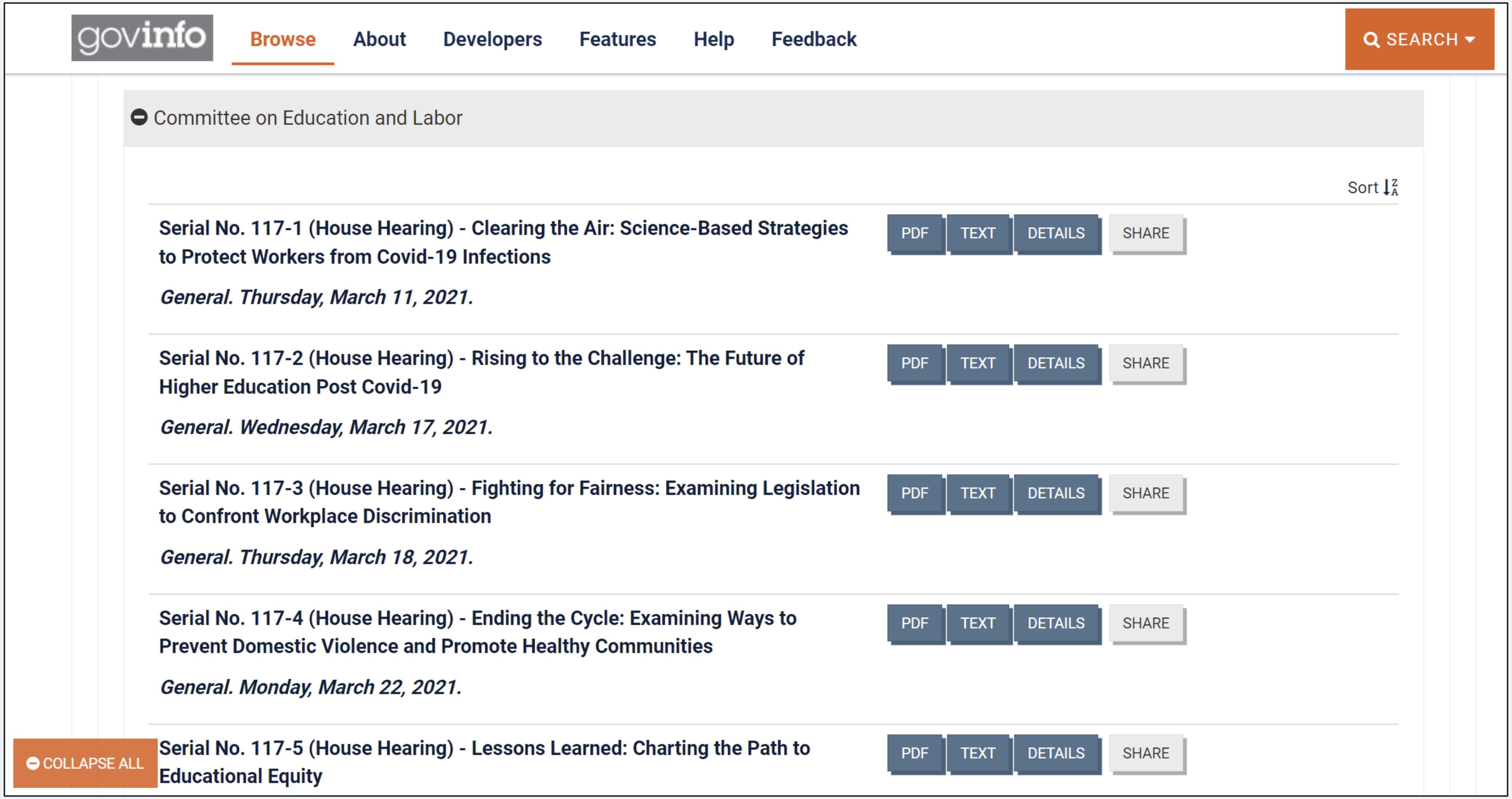
MOVING ON: KEY ELEMENTS OF OUR FULL SCRIPT

1. Navigate the site:
   need to tell our browser where to click
   > Tool: XPath

2. Create our local copy:
   store all pages of interest

3. Make capture comprehensive:
   go through all the pages we need
   > Tool: for loops

3. MAKE CAPTURE COMPREHENSIVE



Committee on Education and Labor

Serial No. 117-1 (House Hearing) - Clearing the Air: Science-Based Strategies to Protect Workers from Covid-19 Infections

General. Thursday, March 11, 2021.

Serial No. 117-2 (House Hearing) - Rising to the Challenge: The Future of Higher Education Post Covid-19

General. Wednesday, March 17, 2021.

Serial No. 117-3 (House Hearing) - Fighting for Fairness: Examining Legislation to Confront Workplace Discrimination

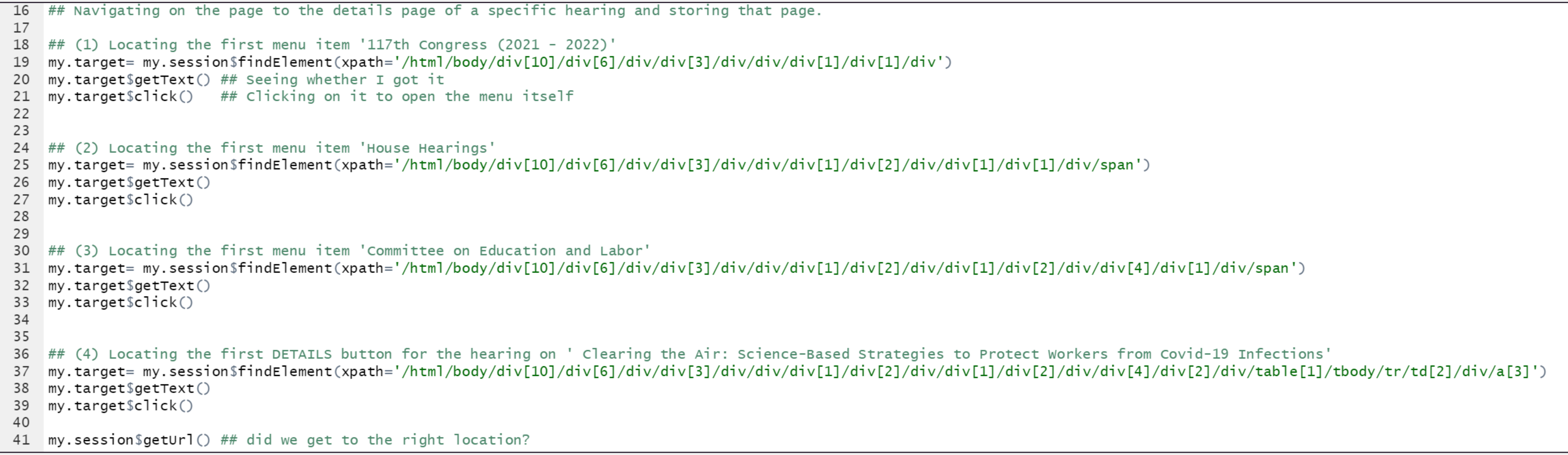
General. Thursday, March 18, 2021.

Serial No. 117-4 (House Hearing) - Ending the Cycle: Examining Ways to Prevent Domestic Violence and Promote Healthy Communities

General. Monday, March 22, 2021.

Serial No. 117-5 (House Hearing) - Lessons Learned: Charting the Path to Educational Equity

CODE GETS REPETITIVE FAST



```
## Navigating on the page to the details page of a specific hearing and storing that page.

## (1) Locating the first menu item '117th Congress (2021 - 2022)'
my.target= my.session$findElement(xpath='/html/body/div[10]/div[6]/div/div[3]/div/div/div[1]/div[1]/div')
my.target$getText() ## Seeing whether I got it
my.target$click()   ## Clicking on it to open the menu itself


## (2) Locating the first menu item 'House Hearings'
my.target= my.session$findElement(xpath='/html/body/div[10]/div[6]/div/div[3]/div/div/div[1]/div[2]/div/div[1]/div[1]/div/span')
my.target$getText()
my.target$click()


## (3) Locating the first menu item 'Committee on Education and Labor'
my.target= my.session$findElement(xpath='/html/body/div[10]/div[6]/div/div[3]/div/div/div[1]/div[2]/div/div[1]/div[2]/div/div[4]/div[1]/div/span')
my.target$getText()
my.target$click()


## (4) Locating the first DETAILS button for the hearing on ' Clearing the Air: Science-Based Strategies to Protect Workers from Covid-19 Infections'
my.target= my.session$findElement(xpath='/html/body/div[10]/div[6]/div/div[3]/div/div/div[1]/div[2]/div/div[1]/div[2]/div/div[4]/div[2]/div/table[1]/tbody/tr/td[2]/div/a[3]')
my.target$getText()
my.target$click()

my.session$getUrl() ## did we get to the right location?
```

1. Move repetitions into a for loop
2. Extract loop elements directly from the HTML

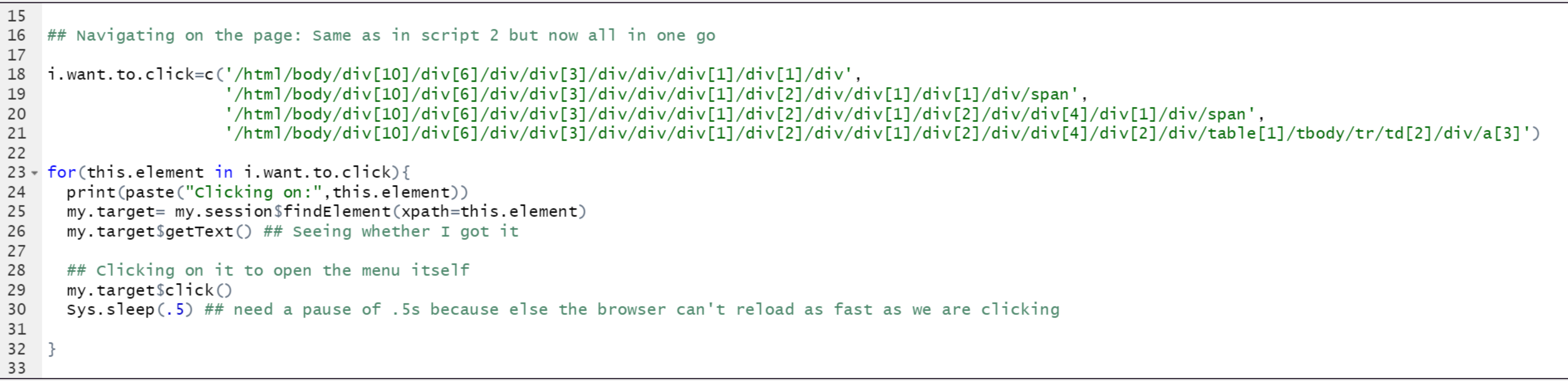
FOR LOOPS
COLLECTING FROM THOUSANDS OF PAGES

- You have to repeat the same operation x times

- for loops avoid code repetition

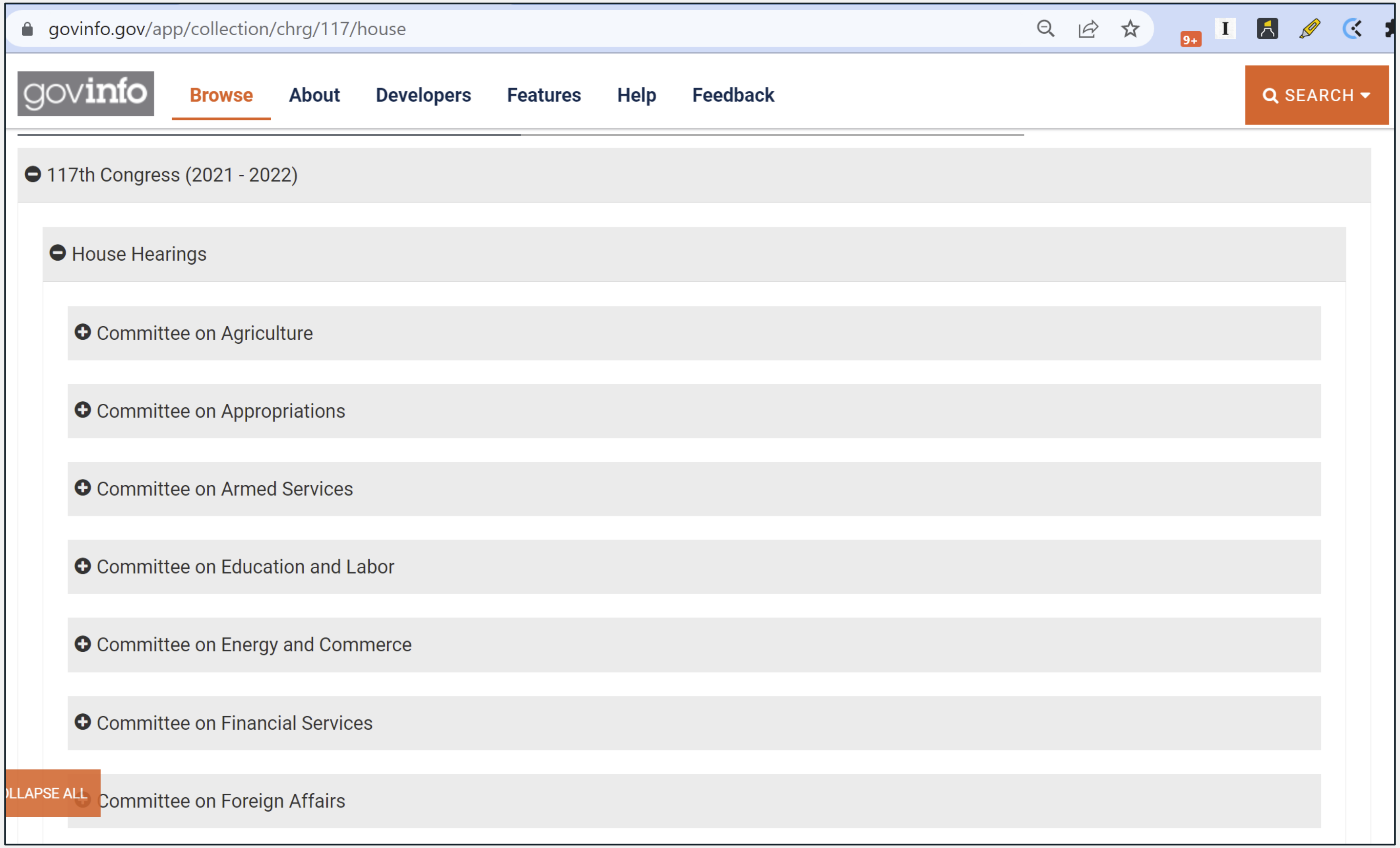
```
for (count in 1:1000) {
  print(count)
}
```

Move repetitions into a for loop



```
## Navigating on the page: Same as in script 2 but now all in one go

i.want.to.click=c('/html/body/div[10]/div[6]/div/div[3]/div/div/div[1]/div[1]/div',
                  '/html/body/div[10]/div[6]/div/div[3]/div/div/div[1]/div[2]/div/div[1]/div[1]/div/span',
                  '/html/body/div[10]/div[6]/div/div[3]/div/div/div[1]/div[2]/div/div[1]/div[2]/div/div[4]/div[1]/div/span',
                  '/html/body/div[10]/div[6]/div/div[3]/div/div/div[1]/div[2]/div/div[1]/div[2]/div/div[4]/div[2]/div/table[1]/tbody/tr/td[2]/div/a[3]')

for(this.element in i.want.to.click){
  print(paste("Clicking on:",this.element))
  my.target= my.session$findElement(xpath=this.element)
  my.target$getText() ## Seeing whether I got it

  ## Clicking on it to open the menu itself
  my.target$click()
  Sys.sleep(.5) ## need a pause of .5s because else the browser can't reload as fast as we are clicking

}

```

Extract loop elements directly from the HTML



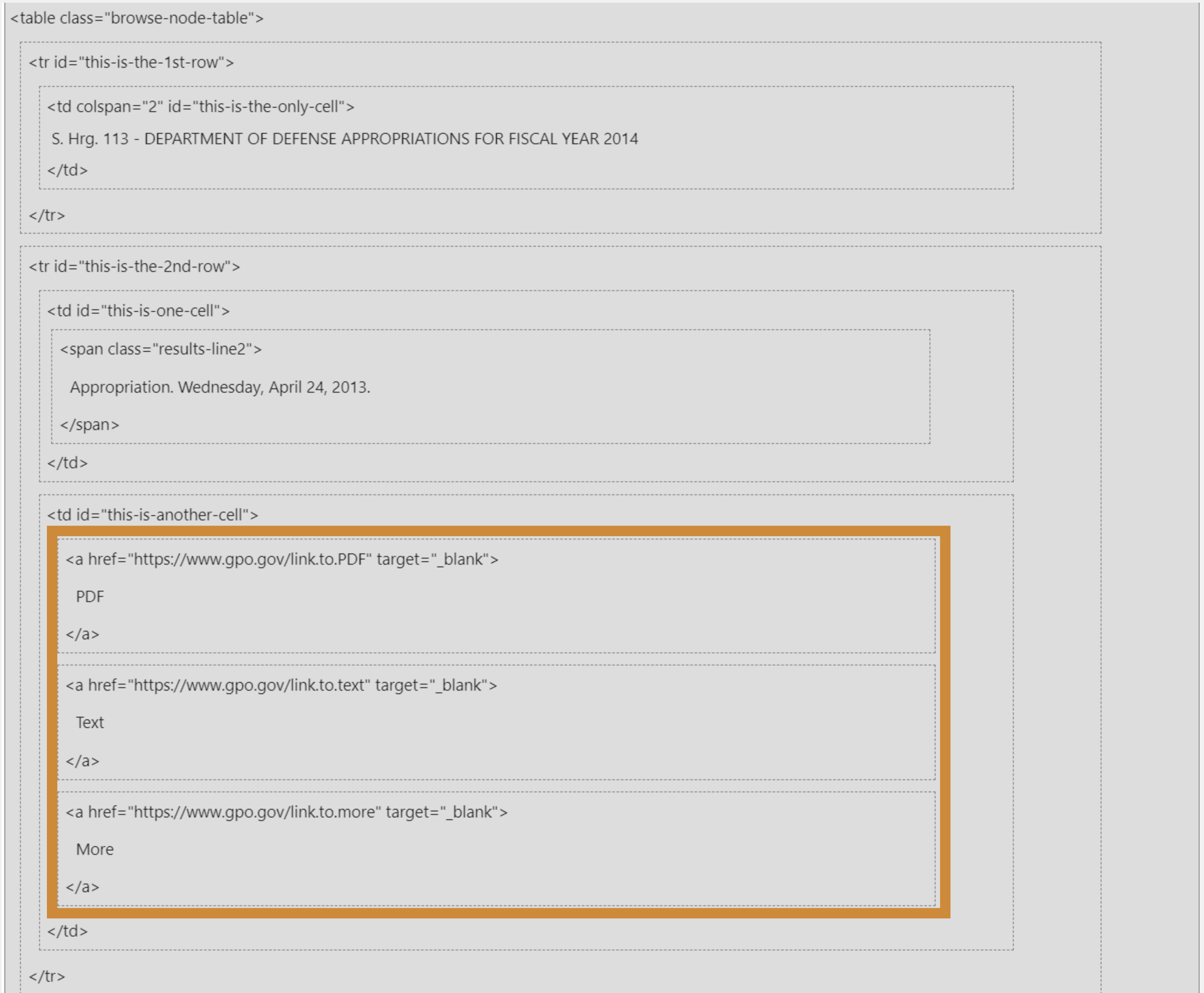
govinfo.gov/app/collection/chrg/117/house

117th Congress (2021 - 2022)

House Hearings

- Committee on Agriculture
- Committee on Appropriations
- Committee on Armed Services
- Committee on Education and Labor
- Committee on Energy and Commerce
- Committee on Financial Services
- Committee on Foreign Affairs

CREATE A LIST OF ALL ELEMENTS IN AN XPATH LOCATION



```
<table class="browse-node-table">
  <tr id="this-is-the-1st-row">
    <td colspan="2" id="this-is-the-only-cell">
     S. Hrg. 113 - DEPARTMENT OF DEFENSE APPROPRIATIONS FOR FISCAL YEAR 2014
    </td>
  </tr>
  <tr id="this-is-the-2nd-row">
    <td id="this-is-one-cell">
      <span class="results-line2">
        Appropriation. Wednesday, April 24, 2013.
      </span>
    </td>
    <td id="this-is-another-cell">
      <a href="https://www.gpo.gov/link.to.PDF" target="_blank">
        PDF
      </a>
      <a href="https://www.gpo.gov/link.to.text" target="_blank">
        Text
      </a>
      <a href="https://www.gpo.gov/link.to.more" target="_blank">
        More
      </a>
    </td>
  </tr>
```

Extract the common root

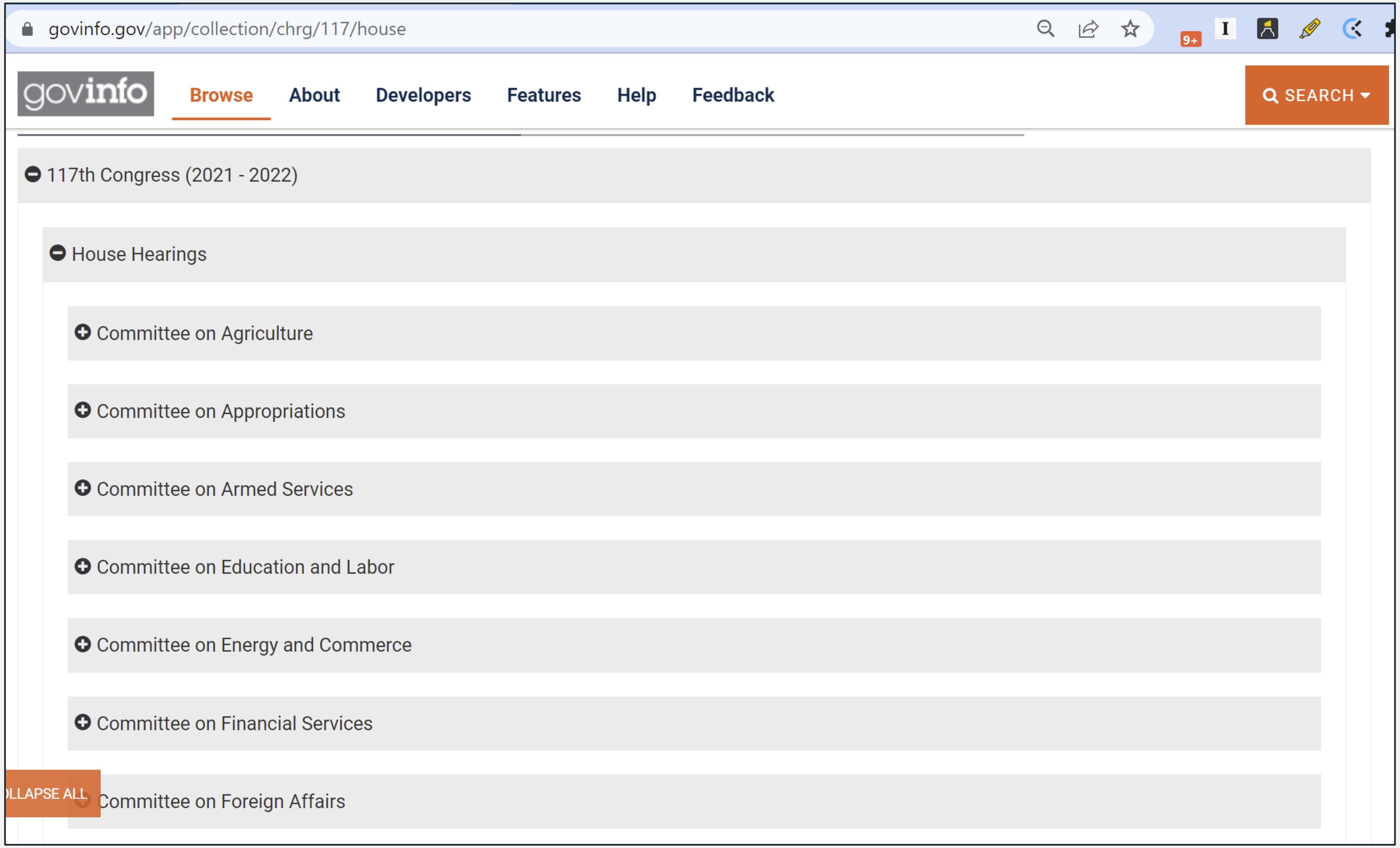
XPaths to all links in this example

- //table/tr[2]/td[2]/a[1]
- //table/tr[2]/td[2]/a[2]
- //table/tr[2]/td[2]/a[3]

common root

- Finding the common root yields 3 elements.

For loop exercise:
Print the names of all House committees into your R console



govinfo.gov/app/collection/chrg/117/house

117th Congress (2021 - 2022)

House Hearings

Committee on Agriculture

Committee on Appropriations

Committee on Armed Services

Committee on Education and Labor

Committee on Energy and Commerce

Committee on Financial Services

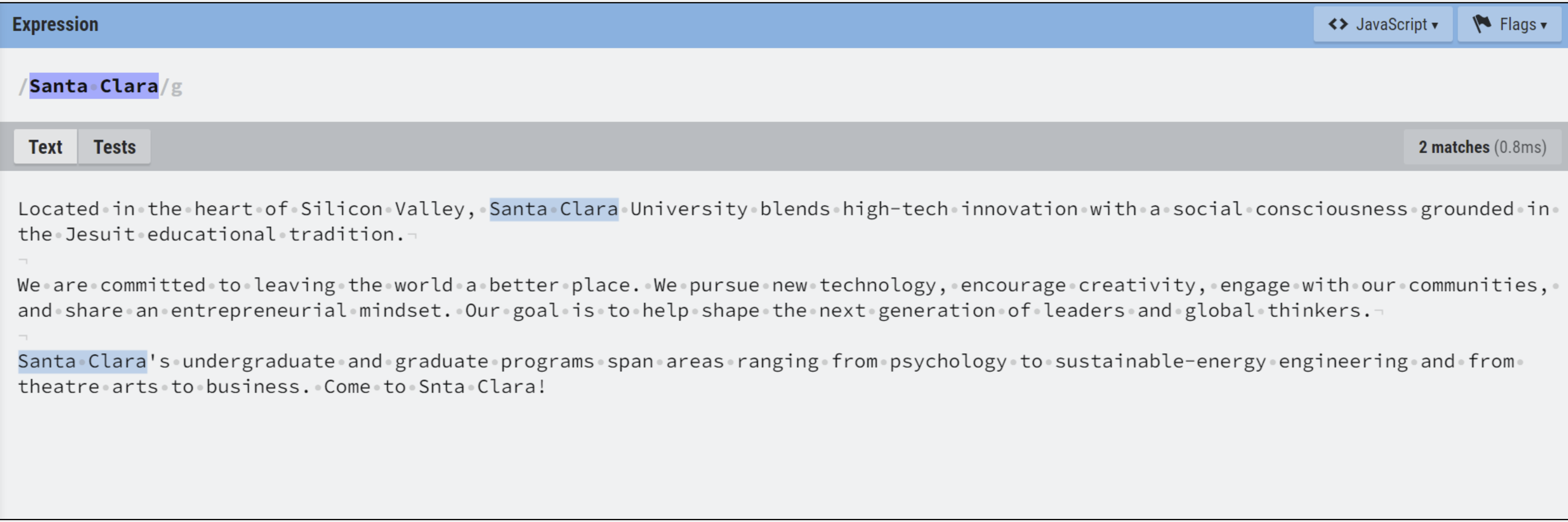
Committee on Foreign Affairs

2.

Parsing the data

Introduction to Regular Expressions.

USUAL TEXT SEARCH IS VERY RIGID



```
/Santa Clara/g
```

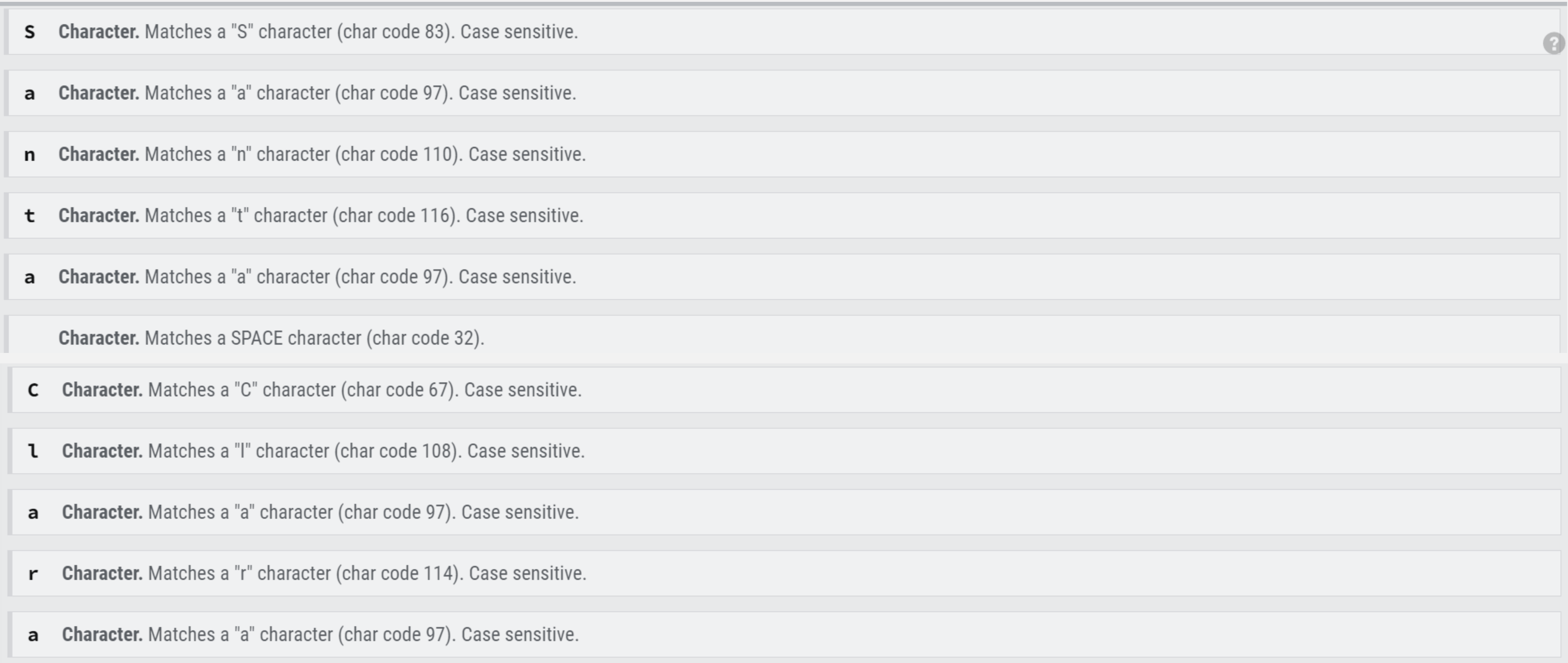
2 matches (0.8ms)

Located in the heart of Silicon Valley, Santa Clara University blends high-tech innovation with a social consciousness grounded in the Jesuit educational tradition.

We are committed to leaving the world a better place. We pursue new technology, encourage creativity, engage with our communities, and share an entrepreneurial mindset. Our goal is to help shape the next generation of leaders and global thinkers.

Santa Clara's undergraduate and graduate programs span areas ranging from psychology to sustainable-energy engineering and from theatre arts to business. Come to Snta Clara!

THIS IS THE QUERY AS RECEIVED BY THE COMPUTER



S **Character.** Matches a "S" character (char code 83). Case sensitive.

a **Character.** Matches a "a" character (char code 97). Case sensitive.

n **Character.** Matches a "n" character (char code 110). Case sensitive.

t **Character.** Matches a "t" character (char code 116). Case sensitive.

a **Character.** Matches a "a" character (char code 97). Case sensitive.

Character. Matches a SPACE character (char code 32).

C **Character.** Matches a "C" character (char code 67). Case sensitive.

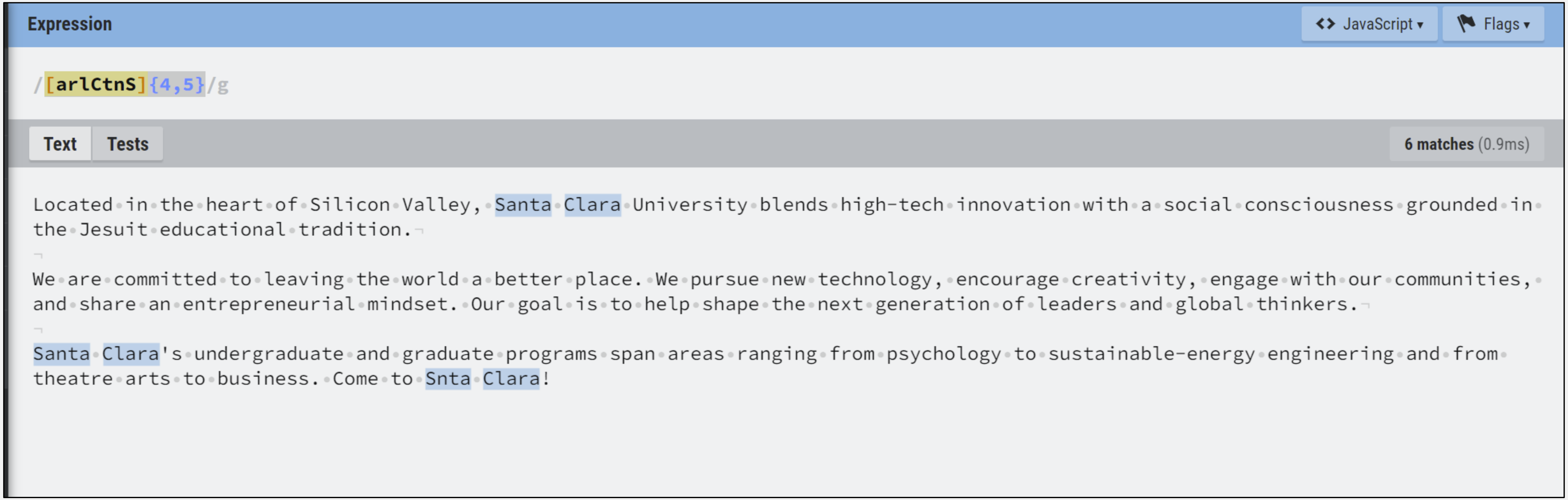
l **Character.** Matches a "l" character (char code 108). Case sensitive.

a **Character.** Matches a "a" character (char code 97). Case sensitive.

r **Character.** Matches a "r" character (char code 114). Case sensitive.

a **Character.** Matches a "a" character (char code 97). Case sensitive.

REGEX FLEXIBILITY HELPS YOU CLEAN UP + EXTRACT FROM TEXT



```
/[arlCtnS]{4,5}/g
```

6 matches (0.9ms)

Located in the heart of Silicon Valley, Santa Clara University blends high-tech innovation with a social consciousness grounded in the Jesuit educational tradition.

We are committed to leaving the world a better place. We pursue new technology, encourage creativity, engage with our communities, and share an entrepreneurial mindset. Our goal is to help shape the next generation of leaders and global thinkers.

Santa Clara's undergraduate and graduate programs span areas ranging from psychology to sustainable-energy engineering and from theatre arts to business. Come to Snta Clara!

REGEX SUPPORTS FLEXIBLE SEARCH

- ► {Santa Clara est. 1851}
- ► Features to describe an expression:
  - > Character types
  - > Character quantifiers
  - > Character location + sequence

ReGex commands

CHARACTER TYPES

- **.** any kind of symbol
- **\d** all or **\D** no digits 0-9
- **\w** all or **\W** no alphanumeric characters (a-z, 0-9)
- **\s** all or **\S** no whitespace (as well as \n \t \v)
- **[**rlCtnaS**]** any of the characters in any order
  - **[**A-Z**]** all capitalized letters
  - **[**a-z**]** all letters in small type
  - **[**0-9**]** all numbers

REGEX COMMANDS

CHARACTER QUANTIFIERS

- Default x=1
- ***** is x=0 or more
- **+** is x= 1 or more
- **{n,m}** is n<= x <=m
- **?** means min(x) incl. zero x
- **()** designates groups

ReGex commands

Character locations, sequence + groups

- **^** means "start of string"
- **$** means "end of string"
- **(?=**abc**)** matches something written before "abc"
- **(?!**abc**)** matches something not written before "abc"
- **(?<=**abc**)** matches something written after "abc"
- **(?<!**abc**)** matches something not written after "abc"

RegEx commands

Avoiding special characters

- What do you do when you want to find a special RegEx character?
  - `+ ? . $` etc.
- Add a backslash (in R: two):
  - `\\+`
  - `\\?`
  - `\\.`
  - `\\$`

REGEX IN R

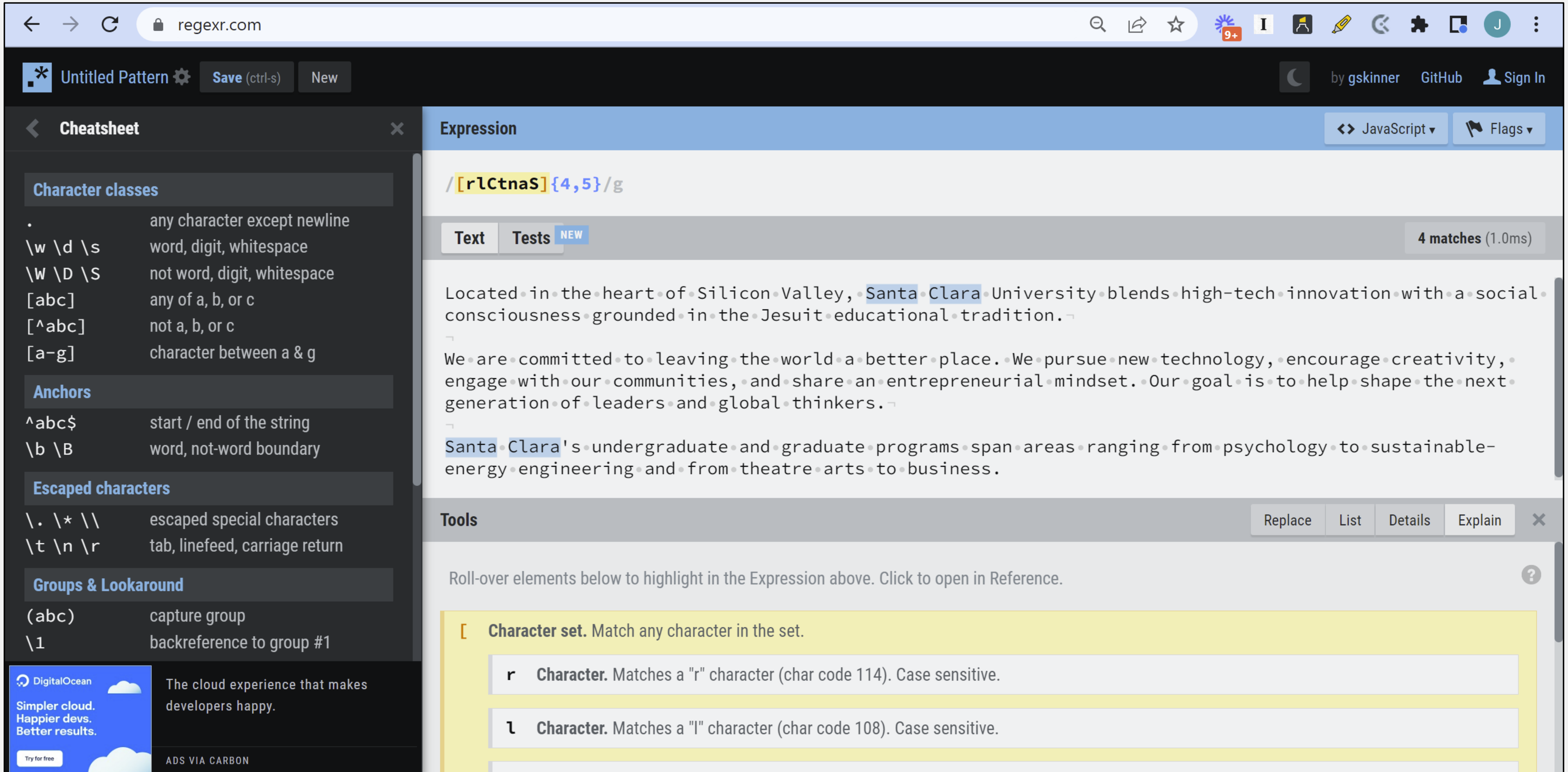
- ► {base} library
  - > Matching:
    - » grepl(pattern, x, ignore.case = FALSE)
    - » grepl(““, ““) →
  - > Substitute:
    - » gsub(pattern, replacement, x, ignore.case = FALSE)
    - » gsub(““, ““, ““) →
- ► {stringr} library
  - > Several additional functions.
    - » Extraction: str_extract_all(string, pattern)
    - » Match count: str_count(string, pattern)

REGEX PRACTICE EXERCISE

- Please open the practice script

Best RegEx playground

REGEXR.COM



Character classes

`.`	any character except newline
`\w \d \s`	word, digit, whitespace
`\W \D \S`	not word, digit, whitespace
`[abc]`	any of a, b, or c
`[^abc]`	not a, b, or c
`[a-g]`	character between a & g

Anchors

`^abc$`	start / end of the string
`\b \B`	word, not-word boundary

Escaped characters

`\. \* \\`	escaped special characters
`\t \n \r`	tab, linefeed, carriage return

Groups & Lookaround

`(abc)`	capture group
`\1`	backreference to group #1

```
/[rlCtnaS]{4,5}/g
```

4 matches (1.0ms)

```
Located in the heart of Silicon Valley, Santa Clara University blends high-tech innovation with a social
consciousness grounded in the Jesuit educational tradition.

We are committed to leaving the world a better place. We pursue new technology, encourage creativity,
engage with our communities, and share an entrepreneurial mindset. Our goal is to help shape the next
generation of leaders and global thinkers.

Santa Clara's undergraduate and graduate programs span areas ranging from psychology to sustainable-
energy engineering and from theatre arts to business.
```

Roll-over elements below to highlight in the Expression above. Click to open in Reference.

[**Character set.** Match any character in the set.

r **Character.** Matches a "r" character (char code 114). Case sensitive.

l **Character.** Matches a "l" character (char code 108). Case sensitive.

RegEx resources

- RegEx sandbox website: https://regexr.com/
- Extensive documentation: http://www.regular-expressions.info
- Flexible text editor: https://www.sublimetext.com/

3. Organizing the data

Basics of relational database design

Single-table setup quickly inefficient

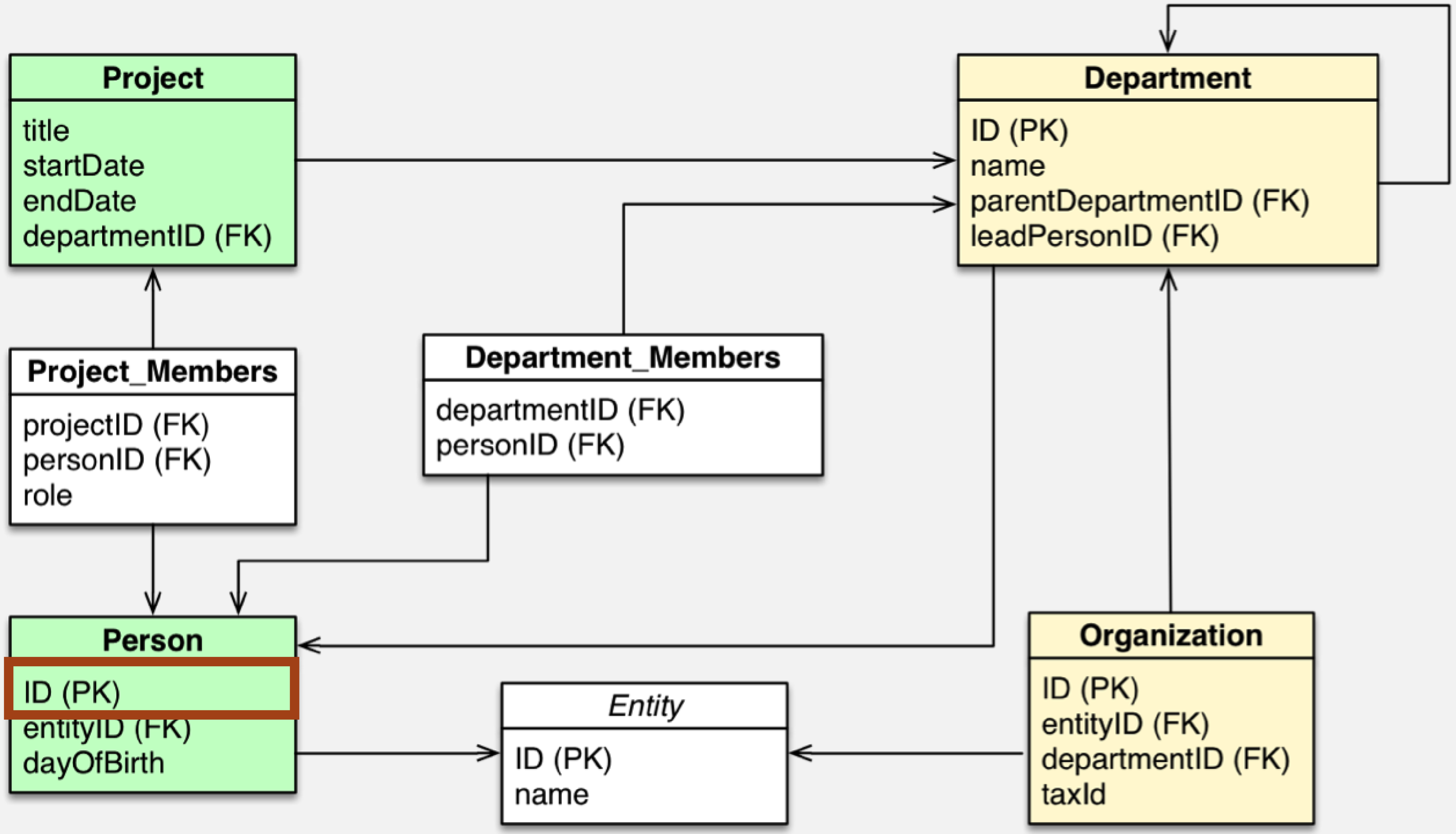
Sources

- ► Repetitions
- ► Unused variables
- ► Large volumes (of text)

Result:

- ► Lots of computational power
- ► Low speed

Basic setup for a list of projects and their members



ID's to organise and merge the data

THAT WAS IT …

1. Harvesting the data:
   Headless **web scraping** using R.

2. Parsing the data:
   Introduction to **Regular Expressions**.

3. Organizing the data:
   Basics of **relational database design**.

All material at:
https://github.com/johannesfritz/scu-harvesting

THANK YOU!

END OF SHORT COURSE