SCRAPING DATA FROM THE WEB I: FREE RESOURCES FOR STORYTELLERS

CHARLES BERRET
BROWN INSTITUTE FOR MEDIA INNOVATION
COLUMBIA UNIVERSITY

DATA STORYTELLING AT BOSTON UNIVERSITY
JUNE 6, 2017

TWO SESSIONS

SCRAPING DATA FROM THE WEB I: FREE RESOURCES FOR STORYTELLERS

BREAK (10:30–10:45)

SCRAPING DATA FROM THE WEB II: APPLYING YOUR NEW SKILLS

SCRAPING DATA FROM THE WEB I: FREE RESOURCES FOR STORYTELLERS

PART 1 – INTRODUCTION TO WEB SCRAPING (9–9:45)

- WHAT IS WEB SCRAPING?
- METHODS AND TOOLS FOR WEB SCRAPING
- HOW SCRAPED DATA CAN BE USED IN STORIES

PART 2 – GUIDED WALKTHROUGH (9:45–10:30)

- HOW TO SCRAPE THE WEB WITH GOOGLE SHEETS

WHAT YOU NEED FOR TODAY'S LESSON

- Modern Web Browser (Firefox, Chrome, Safari)
- Google Sheets (free with Google/GMail account)

WHAT YOU WILL LEARN

- How to move information from a website to a spreadsheet
- How to make scraped data useful

WHAT IS WEB SCRAPING?

- Copying data from a website and storing it in another form so that it's useful to you.

	A	B	C	D	E
1	Member	Trade mil. USD (	Nom. GDP mil. U	PPP GDP mil. U	Nom. GDP per c
2	Argentina	142,370	594,975	924,481	13,589
3	Australia	496,700	1,343,608	1,246,480	50,962
4	Brazil	484,600	2,140,940	3,217,986	8,670
5	Canada	947,200	1,532,340	1,742,656	43,332
6	China	4,201,000	12,361,737	23,066,642	7,990
7	France	1,212,300	2,570,023	2,833,151	37,675
8	Germany	2,666,600	3,618,621	4,122,402	40,997
9	India	650,600	2,457,748	9,585,371	1,617
10	Indonesia	346,100	1,014,867	3,258,727	3,382
11	Italy	948,600	1,895,318	2,289,578	29,887
12	Japan	1,522,400	5,106,259	5,068,064	32,486
13	South Korea	1,170,900	1,521,000	2,029,861	27,195
14	Mexico	613,500	1,124,316	2,410,946	9,009
15	Russia	644,200	1,442,406	3,868,332	9,055
16	Saudi Arabia	521,600	689,004	1,803,419	20,813
17	South Africa	200,100	288,199	758,123	5,695
18	Turkey	417,000	769,474	1,756,510	9,437
19	United Kingdom	1,189,400	2,609,912	2,877,505	43,771
20	United States	3,944,000	19,377,203	19,377,203	55,805
21	European Union	4,485,000	16,970,024	20,745,303	31,918

WHAT IS WEB SCRAPING?

- Copying data from a website and storing it in another form so that it's useful to you.

  - Text
  - Numbers
  - Images
  - Videos

WHAT IS WEB SCRAPING?

- Regularities in HTML allow you to pick out just the elements you want
- Advantage: you get just the data you want, and sometimes it's data that was not previously available in a useful form
- Most scraping requires coding, or else software that handles the code behind the scenes
- No coding in today's exercise, but it will prepare you to start thinking like a programmer

WEB SCRAPING TOOLS

- Commercial software and services (import.io, dex.io, octoparse, parsehub, fminer)
- Browser extensions (OutwitHub for Firefox, Scraper for Chrome)
- Code (Python, Ruby, etc.)

SCRAPING WITH CODE

- Python libraries
  - Beautiful Soup (best for beginners)
  - Scrapy (crawls webpages)
  - Selenium (headless browser)

- Automate data gathering with cron jobs

LIMITS TO SCRAPING

- Poorly formatted HTML
- Authentication systems, paywalls, CAPTCHAs
- Systems that use sessions or cookies to track user activity
- Other access restrictions and usage caps
- Information spread across multiple pages

USING SCRAPED DATA

ProPublica | See more at VITAL SIGNS DONATE

Dollars for Docs

By Charles Ornstein, Lena Groeger, Mike Tigas, and Ryann Grochowski Jones, ProPublica. Updated December 13, 2016

Pharmaceutical and medical device companies are now required by law to release details of their payments to a variety of doctors and U.S. teaching hospitals for promotional talks, research and consulting, among other categories. Use this tool to search for general payments (excluding research and ownership interests) made from August 2013 to December 2015. | Related Story: We've Updated Dollars for Docs. Here's What's New. »

Has Your Doctor Received Drug or Device Company Money?

All States | Search

For example: Andrew Jones, Boston, 10013

$6.25B in disclosed payments | **810,716** doctors | **1,171** teaching hospitals | **1,866** companies

Totals listed below account for all payments from August 2013 to December 2015.

Top 50 Companies

Click on a company to see how its payments break down by drug, device or doctor. Or, see all companies »

Highest-Earning Doctors

NAME	PAYMENTS
ROGER JACKSON Orthopaedic Surgery of the Spine	$64.1M

About the Dollars for Docs Data

Details behind our drug company money database.

Download the Data

The entire data set is available for purchase in the ProPublica Data Store.

Source

The Centers for Medicare and Medicaid Services Open Payments data.

Archive

Search for payments made by 17 drug companies between 2009 and 2013.

Patients, Take Action

We want to know how you've used or

USING SCRAPED DATA – QUESTIONS TO ASK YOURSELF

- Is the material copyrighted?
- Is the dataset already available elsewhere?
- Have you gathered enough data?
- Are the data accurate?
- What can I learn from the data? And how can I tell a story with the data?

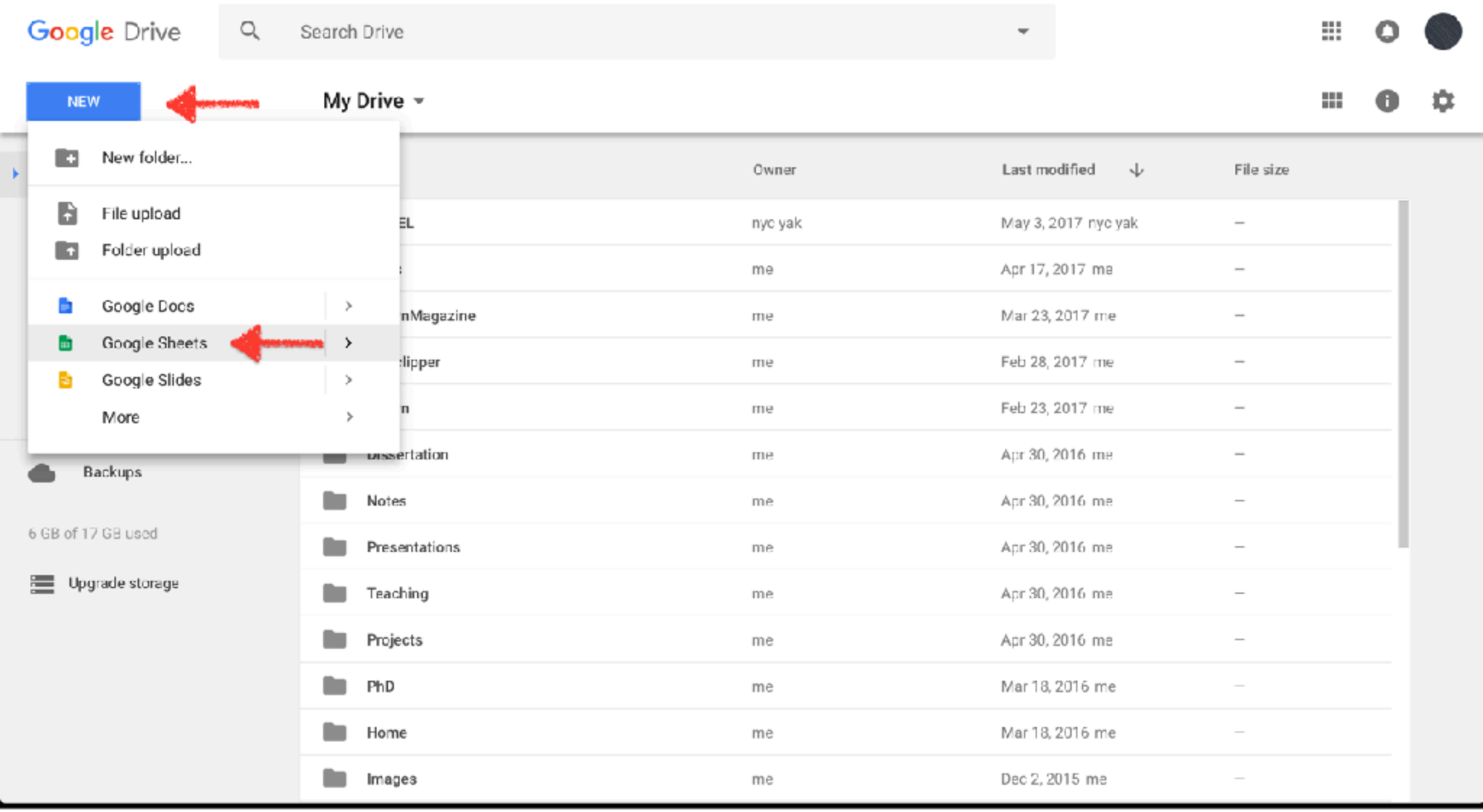
PART II:

SCRAPING THE WEB WITH GOOGLE SHEETS

open your web browser and visit:

drive.google.com

DRIVE.GOOGLE.COM



COPY AND PASTE (SOMETIMES) WORKS — BUT IT'S TEDIOUS

- WIKIPEDIA.ORG —> G20
- HTTPS://EN.WIKIPEDIA.ORG/WIKI/G20

Member country data [edit]

Member	Trade mil. USD (2014)	Nom. GDP mil. USD (2017)[30]	PPP GDP mil. USD (2017)[30]	Nom. GDP per capita USD (2015)[30]	PPP GDP per capita USD (2015)[30]	HDI (2015)	Population (2014)	Area	P5	G4	G7	BRICS	MIKTA	DAC	OECD	C'wth	N11
Argentina	142,370	594,975	924,481	13,589	22,554	0.827	42,961,000	2,780,400	✗	✗	✗	✗	✗	✗	✗	✗	✗
Australia	496,700	1,343,608	1,246,480	50,962	47,389	0.939	23,599,000	7,692,024	✗	✗	✗	✗	✓	✓	✓	✓	✗
Brazil	484,600	2,140,940	3,217,986	8,670	16,155	0.754	202,768,000	8,515,767	✗	✓	✗	✓	✗	✗	✗	✗	✗
Canada	947,200	1,532,340	1,742,656	43,332	44,967	0.920	35,467,000	9,984,670	✗	✗	✓	✗	✗	✓	✓	✓	✗
China	4,201,000	12,361,737	23,066,642	7,990	14,107	0.738	1,367,520,000	9,572,900	✓	✗	✗	✓	✗	✗	✗	✗	✗
European Union	4,485,000	16,970,024	20,745,303	31,918	37,852	0.876	505,570,700	4,422,773	✗	✗	✓	✗	✗	✓	✗	✗	✗
France	1,212,300	2,570,023	2,833,151	37,675	41,181	0.897	63,951,000	640,679	✓	✗	✓	✗	✗	✓	✓	✗	✗
Germany	2,866,600	3,618,621	4,122,402	40,997	46,893	0.926	89,940,000	357,114	✗	✓	✓	✗	✗	✓	✓	✗	✗
India	850,600	2,457,748	9,585,371	1,617	6,162	0.624	1,259,695,000	3,287,263	✗	✓	✗	✓	✗	✗	✗	✓	✗
Indonesia	346,100	1,014,867	3,256,727	3,362	11,126	0.689	25[illegible],490,000	1,904,569	✗	✗	✗	✗	✓	✗	✗	✗	✓
Italy	948,600	1,895,318	2,289,578	29,867	35,708	0.887	60,665,551	301,336	✗	✗	✓	✗	✗	✓	✓	✗	✗
Japan	1,522,400	5,106,259	5,066,064	32,486	38,054	0.903	127,061,000	377,930	✗	✓	✓	✗	✗	✓	✓	✗	✗
Mexico	813,500	1,124,316	2,410,946	9,009	17,534	0.762	119,581,789	1,964,375	✗	✗	✗	✗	✓	✗	✓	✗	✓
Russia	844,200	1,442,406	3,866,332	9,055	25,411	0.804	146,300,000	17,098,242	✓	✗	✗	✓	✗	✗	✗	✗	✗
Saudi Arabia	521,600	689,004	1,803,419	20,813	53,624	0.847	33,624,000	2,149,690	✗	✗	✗	✗	✗	✗	✗	✗	✗
South Africa	200,100	288,199	758,123	5,695	13,165	0.666	53,699,000	1,221,037	✗	✗	✗	✓	✗	✗	✗	✓	✗
South Korea	1,170,900	1,521,000	2,029,861	27,195	36,511	0.901	58,437,000	100,210	✗	✗	✗	✗	✓	✓	✓	✗	✓
Turkey	417,000	769,474	1,756,510	9,437	20,438	0.767	77,324,000	783,562	✗	✗	✗	✗	✓	✗	✓	✗	✓
United Kingdom	1,189,400	2,609,912	2,877,505	43,771	41,159	0.909	64,511,000	242,495	✓	✗	✓	✗	✗	✓	✓	✓	✗
United States	3,944,000	19,377,203	19,377,203	55,805	55,805	0.920	318,523,000	9,526,468	✓	✗	✓	✗	✗	✓	✓	✗	✗

COPY AND PASTE (SOMETIMES) WORKS — BUT IT'S TEDIOUS

- CLICK AND DRAG TO SELECT THE ENTIRE TABLE
- TYPE COMMAND+C OR CTRL+C TO COPY
- GO TO GOOGLE SHEETS AND TYPE COMMAND+V OR CTRL-V TO PASTE

Member country data [edit]

Member	Trade mil. USD (2014)	Nom. GDP mil. USD (2017)[30]	PPP GDP mil. USD (2017)[30]	Nom. GDP per capita USD (2015)[30]	PPP GDP per capita USD (2015)[30]	HDI (2015)	Population (2014)	Area	P5	G4	G7	BRICS	MIKTA	DAC	OECD	C'wth	N11
Argentina	142,370	594,975	924,481	13,589	22,554	0.827	42,961,000	2,780,400	✗	✗	✗	✗	✗	✗	✗	✗	✗
Australia	496,700	1,343,608	1,246,480	50,962	47,389	0.939	23,599,000	7,692,024	✗	✗	✗	✗	✓	✓	✓	✓	✗
Brazil	484,600	2,140,940	3,217,986	8,670	16,155	0.754	202,768,000	8,515,767	✗	✓	✗	✓	✗	✗	✗	✗	✗
Canada	947,200	1,532,340	1,742,656	43,332	44,967	0.920	35,467,000	9,984,670	✗	✗	✓	✗	✗	✓	✓	✓	✗
China	4,201,000	12,361,737	23,066,642	7,990	14,107	0.738	1,367,520,000	9,572,900	✓	✗	✗	✓	✗	✗	✗	✗	✗
European Union	4,485,000	16,970,024	20,745,303	31,918	37,852	0.876	505,570,700	4,422,773	✗	✗	✓	✗	✗	✓	✗	✗	✗
France	1,212,300	2,570,023	2,833,151	37,675	41,181	0.897	63,951,000	640,679	✓	✗	✓	✗	✗	✓	✓	✗	✗
Germany	2,866,600	3,618,621	4,122,402	40,997	46,893	0.926	89,940,000	357,114	✗	✓	✓	✗	✗	✓	✓	✗	✗
India	850,600	2,457,748	9,585,371	1,617	6,162	0.624	1,259,695,000	3,287,263	✗	✓	✗	✓	✗	✗	✗	✓	✗
Indonesia	346,100	1,014,867	3,256,727	3,362	11,126	0.689	25[illegible],490,000	1,904,569	✗	✗	✗	✗	✓	✗	✗	✗	✓
Italy	948,600	1,895,318	2,289,578	29,867	35,708	0.887	60,665,551	301,336	✗	✗	✓	✗	✗	✓	✓	✗	✗
Japan	1,522,400	5,106,259	5,066,064	32,486	38,054	0.903	127,061,000	377,930	✗	✓	✓	✗	✗	✓	✓	✗	✗
Mexico	813,500	1,124,316	2,410,946	9,009	17,534	0.762	119,581,789	1,964,375	✗	✗	✗	✗	✓	✗	✓	✗	✓
Russia	844,200	1,442,406	3,866,332	9,055	25,411	0.804	146,300,000	17,098,242	✓	✗	✗	✓	✗	✗	✗	✗	✗
Saudi Arabia	521,600	689,004	1,803,419	20,813	53,624	0.847	33,624,000	2,149,690	✗	✗	✗	✗	✗	✗	✗	✗	✗
South Africa	200,100	288,199	758,123	5,695	13,165	0.666	53,699,000	1,221,037	✗	✗	✗	✓	✗	✗	✗	✓	✗
South Korea	1,170,900	1,521,000	2,029,861	27,195	36,511	0.901	58,437,000	100,210	✗	✗	✗	✗	✓	✓	✓	✗	✓
Turkey	417,000	769,474	1,756,510	9,437	20,438	0.767	77,324,000	783,562	✗	✗	✗	✗	✓	✗	✓	✗	✓
United Kingdom	1,189,400	2,609,912	2,877,505	43,771	41,159	0.909	64,511,000	242,495	✓	✗	✓	✗	✗	✓	✓	✓	✗
United States	3,944,000	19,377,203	19,377,203	55,805	55,805	0.920	318,523,000	9,526,468	✓	✗	✓	✗	✗	✓	✓	✗	✗

IT WORKED! (SORT OF) BUT IT'S BETTER TO AUTOMATE

Untitled spreadsheet

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

Comments Share

Member

	A	B	C	D	E	F	G	H	I	J	K	L
1	Member	ade mil. USD (201	GDP mil. USD (20	GDP mil. USD (201	P per capita USD (2	P per capita USD (2	HDI (2015)	Population (2014)	Area	P5	G4	G7
2	Argentina	142,370	591,975	924,481	13,589	22,551	0.827	42,961,000	2,780,400			
3	Australia	496,700	1,343,608	1,246,480	50,962	47,380	0.939	23,590,000	7,692,024			
4	Brazil	484,600	2,140,940	3,217,986	8,670	16,155	0.754	202,768,000	8,515,767			
5	Canada	947,200	1,532,340	1,742,856	43,332	44,967	0.92	35,467,000	9,984,670			
6	China	4,201,000	12,361,737	23,066,542	7,990	14,107	0.738	1,357,520,000	9,572,900			
7	European Union	4,485,000	16,970,024	20,745,303	31,918	37,852	0.876	505,570,700	4,422,773			
8	France	1,212,300	2,570,023	2,833,151	37,675	41,181	0.897	63,951,000	640,679			
9	Germany	2,066,600	3,610,621	4,122,402	40,997	46,893	0.926	80,940,000	357,114			
10	India	850,600	2,457,748	9,585,371	1,617	6,162	0.624	1,259,695,000	3,287,263			
11	Indonesia	346,100	1,014,867	3,256,727	3,362	11,125	0.689	251,490,000	1,904,569			
12	Italy	948,600	1,895,318	2,289,578	29,867	35,708	0.837	60,665,551	301,336			
13	Japan	1,522,400	5,106,259	5,066,064	32,486	38,054	0.903	127,061,000	377,930			
14	Mexico	613,500	1,124,316	2,410,946	9,009	17,534	0.762	119,581,709	1,964,375			
15	Russia	844,200	1,442,406	3,866,332	9,055	25,411	0.804	146,300,000	17,098,242			
16	Saudi Arabia	521,600	689,004	1,803,419	20,613	53,624	0.847	30,624,000	2,149,690			
17	South Africa	200,100	288,199	758,123	5,695	13,165	0.666	53,699,000	1,221,037			
18	South Korea	1,170,900	1,521,000	2,029,861	27,195	36,511	0.901	50,437,000	100,210			
19	Turkey	417,000	769,474	1,756,510	9,437	20,430	0.767	77,024,000	783,562			
20	United Kingdom	1,189,400	2,609,912	2,877,505	43,771	41,159	0.909	64,511,000	242,495			
21	United States	3,944,000	19,377,203	19,377,203	55,805	55,805	0.92	318,523,000	9,526,468			
22												
23												
24												

G20 Nations Sheet2

SCRAPING TABLES TO GOOGLE SHEETS

Command to scrape a table from a website:

=ImportHTML(“http://example.com”,“table”,1)

- =ImportHTML(“http://website.com”,“table”,1) **Command** (1,2,3)
- =ImportHTML(“http://website.com”,“table”,1) **“URL”**
- =ImportHTML(“http://website.com”,“table”,1) **“Element Type”**
- =ImportHTML(“http://website.com”,“table”,1) **Element Number**

SCRAPING TABLES TO GOOGLE SHEETS

To scrape a table from https://en.wikipedia.org/wiki/G20

Type this command directly into a spreadsheet box:

=ImportHTML(“https://en.wikipedia.org/wiki/G20”,“table”,1)

- =ImportHTML(“https://en.wikipedia.org/wiki/G20”,“table”,1)
- =ImportHTML(“https://en.wikipedia.org/wiki/G20”,“table”,1)
- =ImportHTML(“https://en.wikipedia.org/wiki/G20”,“table”,1)
- =ImportHTML(“https://en.wikipedia.org/wiki/G20”,“table”,1)

Untitled spreadsheet

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

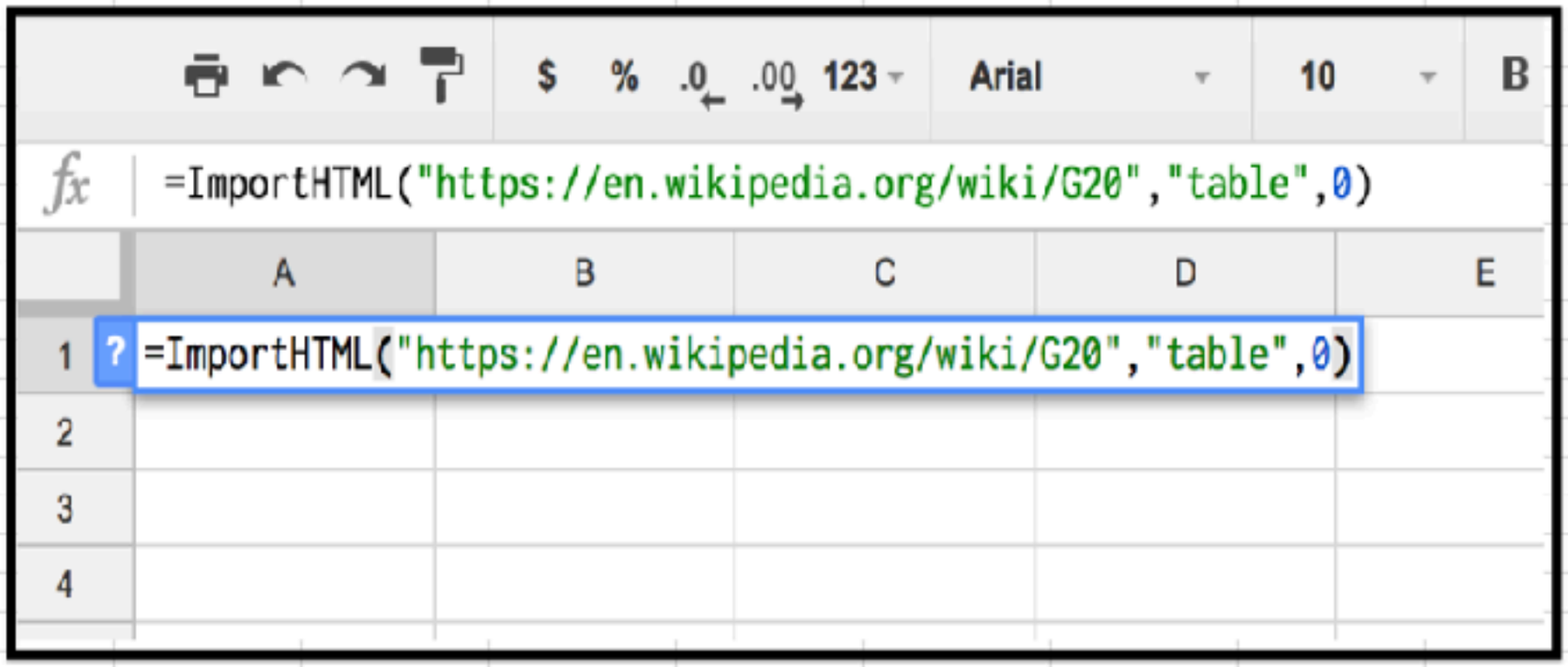
Comments Share

Arial 10

=ImportHTML("https://en.wikipedia.org/wiki/G20","table",0)

=ImportHTML("https://en.wikipedia.org/wiki/G20","table",0)

G20 Nations Sheet2



Untitled spreadsheet

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

Comments Share

=ImportHTML("https://en.wikipedia.org/wiki/G20","table",0)

	A	B	C	D	E	F	G	H	I	J	K	L	
1	Member	Trade mil. USD (	Nom. GDP mil. U	PPP GDP mil. U	Nom. GDP per c	PPP GDP per ca	HDI (2015)	Population (2014	Area	P5	G4	G7	B
2	Argentina	142,370	594,975	924,481	13,589	22,554	0.827	42,961,000	2,780,400	N	N	N	N
3	Australia	496,700	1,343,608	1,246,480	50,962	47,389	0.939	23,599,000	7,692,024	N	N	N	N
4	Brazil	484,600	2,140,940	3,217,986	8,670	16,155	0.754	202,768,000	8,515,767	N	Y	N	Y
5	Canada	947,200	1,532,340	1,742,656	43,332	44,967	0.92	35,467,000	9,984,670	N	N	Y	N
6	China	4,201,000	12,361,737	23,066,642	7,990	14,107	0.738	1,367,520,000	9,572,900	Y	N	N	Y
7	France	1,212,300	2,570,023	2,833,151	37,675	41,181	0.897	63,951,000	640,679	Y	N	Y	N
8	Germany	2,886,600	3,618,621	4,122,402	40,997	46,893	0.926	80,940,000	357,114	N	Y	Y	N
9	India	850,600	2,457,748	9,585,371	1,617	6,162	0.624	1,259,695,000	3,287,263	N	Y	N	Y
10	Indonesia	346,100	1,014,867	3,256,727	3,362	11,126	0.689	251,490,000	1,904,569	N	N	N	N
11	Italy	948,600	1,895,318	2,289,578	29,867	35,708	0.887	60,665,551	301,336	N	N	Y	N
12	Japan	1,522,400	5,106,259	5,086,064	32,486	38,054	0.903	127,061,000	377,930	N	Y	Y	N
13	South Korea	1,170,900	1,521,000	2,029,861	27,195	36,511	0.901	50,437,000	100,210	N	N	N	N
14	Mexico	813,500	1,124,316	2,410,946	9,009	17,534	0.762	119,581,789	1,964,375	N	N	N	N
15	Russia	844,200	1,442,406	3,886,332	9,055	25,411	0.804	146,300,000	17,098,242	Y	N	N	Y
16	Saudi Arabia	521,600	689,004	1,803,419	20,813	53,624	0.847	30,624,000	2,149,690	N	N	N	N
17	South Africa	200,100	288,199	758,123	5,695	13,165	0.666	53,699,000	1,221,037	N	N	N	Y
18	Turkey	417,000	769,474	1,756,510	9,437	20,438	0.767	77,324,000	783,562	N	N	N	N
19	United Kingdom	1,189,400	2,609,912	2,877,505	43,771	41,159	0.909	64,511,000	242,495	Y	N	Y	N
20	United States	3,944,000	19,377,203	19,377,203	55,805	55,805	0.92	318,523,000	9,526,468	Y	N	Y	N
21	European Union	4,485,000	16,970,024	20,745,303	31,918	37,852	0.876	505,570,700	4,422,773	N	N	Y	N
22													
23													
24													

G20 Nations Sheet2

Explore

finding “tables” in the HTML source code of the page

Secure https://en.wikipedia.org/wiki/G20

memberships or other international organisations, such as the G7 and BRICS. Total GDP figures are given in millions of US dollars.

table.wikitable.sortable.jquery-tablesorter 1040×683

Member	Leader position	Head of government	Finance portfolio	Finance minister	Central bank governor
Argentina	President	Mauricio Macri	Minister of Public Finances Minister of the Treasury	Luis Caputo Nicolás Dujovne	Federico Sturzenegger
Australia	Prime Minister	Malcolm Turnbull	Treasurer	Scott Morrison	Philip Lowe
Brazil	President	Michel Temer	Minister of Finance	Henrique Meirelles	Ilan Goldfajn
Canada	Prime Minister	Justin Trudeau	Minister of Finance	Bill Morneau	Stephen Poloz
China	President[note 1]	Xi Jinping[note 1]	Minister of Finance	Xiao Jie	Zhou Xiaochuan
France	President	Emmanuel Macron	Minister of the Economy	Bruno Le Maire	François Villeroy de Galhau
Germany	Chancellor	Angela Merkel	Minister of Finance	Wolfgang Schäuble	Jens Weidmann
India	Prime Minister	Narendra Modi	Minister of Finance	Arun Jaitley	Urjit Patel

Elements Console Sources Network Performance Memory Application Security Audits HTTPS Everywhere

```
▼<table class="wikitable sortable jquery-tablesorter" style="width:100%;"> == $0
  ▼<thead>
    ▼<tr>
      <th class="headerSort" tabindex="0" role="columnheader button" title="Sort ascending">Member</th>
      <th class="headerSort" tabindex="0" role="columnheader button" title="Sort ascending">Leader position</th>
      <th class="headerSort" tabindex="0" role="columnheader button" title="Sort ascending">Head of government</th>
      <th class="headerSort" tabindex="0" role="columnheader button" title="Sort ascending">Finance portfolio</th>
      <th class="headerSort" tabindex="0" role="columnheader button" title="Sort ascending">Finance minister</th>
      <th class="headerSort" tabindex="0" role="columnheader button" title="Sort ascending">Central bank governor</th>
    </tr>
  </thead>
  ▶<tbody>…</tbody>
  <tfoot></tfoot>
</table>
▶<h3>…</h3>
▶<table class="wikitable sortable jquery-tablesorter" style="font-size:90%; width:100%;">…</table>
▶<p>…</p>
▶<p>…</p>
```

html body #content #bodyContent #mw-content-text div.mw-parser-output table.wikitable.sortable.jquery-tablesorter thead tr th.headerSort

Styles Computed Event Listeners »

Filter :hov .cls +

```
element.style {
  width: 100%;
}
table.wikitable load.php?debug=…tor.desktop…:1
{
  margin: ▶1em 0;
  background-color: #f8f9fa;
  border: ▶1px solid #a2a9b1;
  border-collapse: collapse;
  color: #000;
}
table {  load.php?debug=…tor.desktop…:1
  font-size: 100%;
}
table {  user agent stylesheet
```

SCRAPING LISTS TO GOOGLE SHEETS

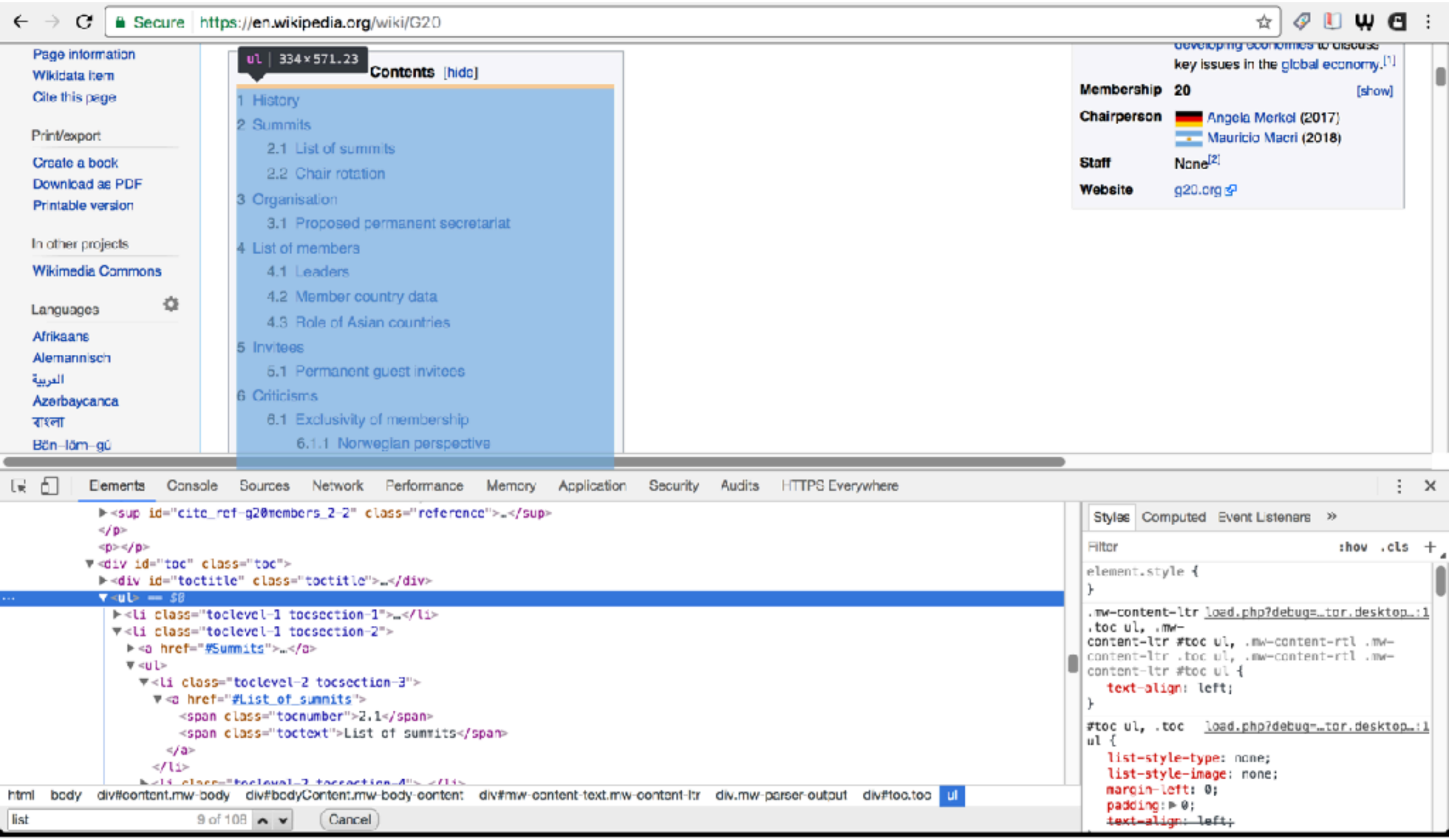
To scrape a list from https://en.wikipedia.org/wiki/G20

Enter the last command, changing “table” to “list” and 1 to 2:

=ImportHTML(“https://en.wikipedia.org/wiki/G20”,“list”,2)

- =ImportHTML(“https://en.wikipedia.org/wiki/G20”,“list”,2)
- =ImportHTML(“https://en.wikipedia.org/wiki/G20”,“list”,2)
- =ImportHTML(“https://en.wikipedia.org/wiki/G20”,“list”,2)
- =ImportHTML(“https://en.wikipedia.org/wiki/G20”,“list”,2)

finding “lists” in the HTML source code of the page



MORE POWERFUL SCRAPING COMMANDS WITH GSHEETS

type command directly into a spreadsheet box:

=IMPORTFEED("https://example.com/whatever")

- =IMPORTFEED(“https://example.com”) **Command** (1)
- =IMPORTFEED(“https://example.com”) **“URL”**

Note: the IMPORTFEED command takes only **ONE** argument

SCRAPING FROM AN RSS FEED WITH GSHEETS

type command directly into a spreadsheet box:

=IMPORTFEED("https://example.com/whatever")

Try one of these RSS feeds from the Library of Congress:

- https://www.congress.gov/rss

Secure https://www.congress.gov/rss

RSS and Email Alerts

Keeping up with Congress is easy with RSS and Email Alerts from Congress.gov. This page allows you to subscribe to a variety of RSS and Email alerts related to Congressional activity and legislation.

Jump to: How to Get Email Alerts

Subscribe to RSS

RSS (Really Simple Syndication) is a technology that delivers news to a computer or mobile device. Congress.gov offers several RSS feeds for use in an RSS reader or RSS-enabled Web browser. For details about RSS, see the RSS help page on loc.gov.

Most-Viewed Bills
A weekly top-ten list of the bills measured by page views on Congress.gov.

Search Tips
Tips on how to perform effective, advanced searches on Congress.gov.

Appropriations Tables
Congress.gov includes appropriation tables to simplify the process of tracking legislative appropriation bills. Each time the table is updated, this feed will let you know what has been added to the table.

Bills Presented to the President
When a piece of legislation that requires the President's signature to become law passes Congress, this feed will alert you that the legislation has been submitted to the President for his or her signature.

On the House Floor Today
Legislation brought to the floor of the United States House of Representatives.

On the Senate Floor Today
Legislation brought to the floor of the United States Senate.

***In Custodia Legis*: Law Librarians of Congress**
The latest posts from the award winning Law Library of Congress blog, *in Custodia Legis*, featuring news about Congress.gov, topical articles concerning foreign and domestic law, and an insider's view of the happenings in and around the Library of Congress and Capitol Hill.

Help

Saved Search Help

Alerts Help

Creating and Using Email Alerts

↑ Back to top

fx =IMPORTFEED("https://www.congress.gov/rss/house-floor-today.xml")

	A	B	C
1	S.1083	https://www.congress.gov/bill/115	A bill to amend section 1214 of title 5, United States Code, to provide for stays during a period that the Merit Systems Protection Board lacks a quorum. (05/25/2017 legislative day)
2	H.R.1973	https://www.congress.gov/bill/115	Protecting Young Victims from Sexual Abuse Act of 2017 (05/25/2017 legislative day)
3	H.R.1761	https://www.congress.gov/bill/115	Protecting Against Child Exploitation Act of 2017 (05/25/2017 legislative day)
4	S.Con.Res.14	https://www.congress.gov/bill/115	A concurrent resolution authorizing the use of Emancipation Hall in the Capitol Visitor Center for an event to celebrate the birthday of King Kamehameha I. (05/24/2017 legislative day)
5	H.Res.350	https://www.congress.gov/bill/115	Permitting official photographs of the House of Representatives to be taken while the House is in actual session on a date designated by the Speaker. (05/24/2017 legislative day)
6	H.R.953	https://www.congress.gov/bill/116	Reducing Regulatory Burdens Act of 2017 (05/24/2017 legislative day)
7	H.R.624	https://www.congress.gov/bill/115	To restrict the inclusion of social security account numbers on Federal documents sent by mail, and for other purposes. (05/24/2017 legislative day)
8	H.R.1293	https://www.congress.gov/bill/115	To amend title 5, United States Code, to require that the Office of Personnel Management submit an annual report to Congress relating to the use of official time by Federal employees. (05/24/2017 legislative day)
9	H.Res.352	https://www.congress.gov/bill/115	Providing for consideration of the bill (H.R. 1973) to prevent the sexual abuse of minors and amateur athletes by requiring the prompt reporting of sexual abuse to law enforcement authorities, and for other purposes; providing for consideration of the bill (H.R. 1761) to amend title 18, United States Code, to criminalize the knowing consent of the visual depiction, or live transmission, of a minor engaged in sexually explicit conduct, and for other purposes; and providing for proceedings during the period from May 26, 2017, through June 5, 2017. (05/24/2017 legislative day)
10	H.R.2473	https://www.congress.gov/bill/115	Enforcing Justice for Victims of Trafficking Act of 2017 (05/23/2017 legislative day)

Sheet1 Sheet2 Sheet4 Sheet3 Sheet5

MORE POWERFUL SCRAPING COMMANDS FOR GSHEETS

Pull data from an API (but don't try this just yet):

=IMPORTDATA("https://data.cityofboston.gov/resource/427a-3cn5.json")

- =IMPORTXML("https://example.com","//a/@href") **Command** (1)
- =IMPORTXML("https://data.cityofboston.gov/resource/427a-3cn5.json") "**URL**"

Untitled spreadsheet

File Edit View Insert Format Data Tools Add-ons Help All changes saved in Drive

charles.bernet@gmail.com

Comments Share

Inconsolata 10

=IMPORTDATA("https://data.cityofboston.gov/resource/427a-3cn5.json")

	A	B	C	D	E	F	G	H	I	
1	[{":@computed_	address:"1000 Washington ST"	businessname:"1000 Wash	city:"Roxbury"	comments:"Prep	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalowner:"TER	licensec
2		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	comments:"Clear	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalow
3		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	comments:"clear	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalow
4		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	comments:"Provi	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalow
5		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalowner:"TER	licensec
6		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	comments:"clear	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalow
7		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	comments:"Provi	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalow
8		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	comments:"Provi	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalow
9		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	comments:"Keep	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalow
10		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	comments:"Keep	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalow
11		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalowner:"TER	licensec
12		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	comments:"Provi	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalow
13		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	comments:"Cove	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalow
14		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	comments:"Keep	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalow
15		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	comments:"Clear	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalow
16		{":@computed_region_aywg_kpfl	address:"1000 Washingto	businessname:"1000 Washington C	city:"Roxbury"	comments:"Cove	descript:"Eating	expdttm:"2011-1	issdttm:"2011-11-	legalow
17		{":@computed_region_aywg_kpfl	address:"635 Hyde Park A	businessname:"100 Percent Delicia	city:"Roslindale"	comments:"Low	descript:"Eating	expdttm:"2017-1	issdttm:"2013-04	legalow
18		{":@computed_region_aywg_kpfl	address:"635 Hyde Park A	businessname:"100 Percent Delicia	city:"Roslindale"	comments:"PIC	descript:"Eating	expdttm:"2017-1	issdttm:"2013-04	legalow
19		{":@computed_region_aywg_kpfl	address:"635 Hyde Park A	businessname:"100 Percent Delicia	city:"Roslindale"	comments:"Repa	descript:"Eating	expdttm:"2017-1	issdttm:"2013-04	legalow
20		{":@computed_region_aywg_kpfl	address:"635 Hyde Park A	businessname:"100 Percent Delicia	city:"Roslindale"	comments:"Paint	descript:"Eating	expdttm:"2017-1	issdttm:"2013-04	legalow
21		{":@computed_region_aywg_kpfl	address:"635 Hyde Park A	businessname:"100 Percent Delicia	city:"Roslindale"	comments:"Certi	descript:"Eating	expdttm:"2017-1	issdttm:"2013-04	legalow
22		{":@computed_region_aywg_kpfl	address:"635 Hyde Park A	businessname:"100 Percent Delicia	city:"Roslindale"	comments:"Clear	descript:"Eating	expdttm:"2017-1	issdttm:"2013-04	legalow
23		{":@computed_region_aywg_kpfl	address:"635 Hyde Park A	businessname:"100 Percent Delicia	city:"Roslindale"	comments:"All fo	descript:"Eating	expdttm:"2017-1	issdttm:"2013-04	legalow
24		{":@computed_region_aywg_kpfl	address:"635 Hyde Park A	businessname:"100 Percent Delicia	city:"Roslindale"	comments:"Base	descript:"Eating	expdttm:"2017-1	issdttm:"2013-04	legalow

Sheet1 Sheet2 Sheet4 Sheet3 Sheet5

MORE POWERFUL SCRAPING COMMANDS FOR GSHEETS

Scrape all links on a webpage:

=IMPORTXML("https://example.com","//a/@href")

- =IMPORTXML(“https://example.com”,"//a/@href") **Command** (1,2)
- =IMPORTXML(“https://example.com”,”//a/@href") **“URL”**
- =IMPORTXML(“https://example.com”,”//a/@href”) **“Query”** (for links)

FIFTEEN MINUTE BREAK 10:30-45

SCRAPING DATA FROM THE WEB II: APPLYING YOUR NEW SKILLS

CHARLES BERRET
BROWN INSTITUTE FOR MEDIA INNOVATION
COLUMBIA UNIVERSITY

DATA STORYTELLING AT BOSTON UNIVERSITY
JUNE 6, 2017

FOUR WAYS TO SCRAPE WITH GOOGLE SHEETS

- Tables (=IMPORTHTML)
- Lists (=IMPORTHTML)
- RSS Feeds and APIs (=IMPORTFEED)
- Advanced (=IMPORTXML)

EXERCISE: SCRAPE DATA, COLLECT SPREADSHEETS, FIND STORIES

- Choose a topic or question that interests you
- Find a website with information on that subject
- Use Google Sheets commands to scrape data into separate tabs of the spreadsheet
- Gather a personal collection of scraped data that you can study to find patterns, regularities, outliers, or other data insights.