Data Fundamentals Lab: Scraping Data from the Web

In this lab, you will learn how to scrape data from websites using browser extensions and Google Spreadsheets.

Scraping Data Using Browser Extensions

Browser Extensions or Browser Add-Ons are applications that run in your internet browser that enable you to:

- Select data tables on a website, including specific rows and columns
- Copy data from these tables and use it in a spreadsheet application like Excel or Google Sheets

Here is an example, let's use a browser extension to scrape data about Tanzania from a United Nation's website. You can use any of these procedures – depending on the browser you are using: Google Chrome or Mozilla Firefox.

Scraping Data using Browser Extension for Google Chrome

1. Open your Google Chrome Browser.

- 2. To download Scraper Extension, go to <u>https://chrome.google.com/webstore/detail/scraper/mbigbapnjcgaffo</u> <u>hmbkdlecaccepngjd?hl=en</u>
- **3**. In the 'Scraper' window that opens, click the blue **+ADD TO CHROME** button.

OVERVIEW	1	REVIEWS	RE	LATED		
1000	Breaner - Comm	arison of JavaResint fi	amawada - Willingto the free of	neveloparla	Desi	reper
4 Comparison of Jas	aScript frameworks -	frameses	ek Version Si compand	za License Source	e language	· =
		1 Ample SD	K 0.9.3 1 Jul 2010 Variable. Cor Iminited & p	e size: 40 k8 MIT & CPL ? zipped)		Scraper gets data out of web pages
NPuth 8 Obligation	teles (1) in our where (1) is the	2 Angular5	1.2.14 1 Mar 36 kB (min) 2014 compressed)	ed & MIT JavaSorij	ot	and into spreadsheets.
@ Mah	Reference	Gubandon	nd) 0.2 june 2012 20 k8 primiñ	ed) Mit JavaSon	pt	Scraper is a very simple (but limited) data
Column N		5 Data	1.10.3.08 Dec Variable Res	Commercial[1]	Source language +	mining extension for facilitating online research when you need to get data into
1 (1) 6	ramework • 0		2014 Sminifed & p k6 (minifed) (uncompress	mpped, 155 598 kB ed(2)		spreadsheet form quickly. It is intended a
1 12) W	ersion compared · · ·	6 6cho3	3.0.xc1 24 Mar 7 2011	MPL_LCPL or JavaSori DPL Java	pt and/or ?	an easy-to-use tool for intermediate to
1 14 U	icense o C	 7 Ember.js 	1.7.0 19 Aug 95 kB (mmf) 2014 pripted), 34 (mmfred), 1.	ell & MIT JavaSoly 0 kB 5 MB edu	of spt	XPath.
Presett Reset	Scrape			Copy to clipboard Deport to Ge	oogie Doca	++7
Dewnload as PDF	DHTMLX	4.Jun 2014	Variable	GPL & Commercial ⁽¹⁾	JavaScript	- feature: copy data to clipboard (as tab
Printicle version			Variable.			separated values)
,≁Edt inna	Dojo	1.10.3 08 Dec 2014	41 kB (minified & gzipped),	BSD & AFL	JavaScript + HTML	- fix: upgraded oauth for Google Docs
		1.0.1	596 kB (uncompressed) ^[2]			Report Abuse
	Echo3	3.0.rc1 24 Mar 2011	7	MPL, LGPL or GPL	JavaScript and/or Java	• Report Abuse
	Ember.js	1.7.0 19 Aug 2014	95 kB (minified & gzipped), 340 kB (minified), 1.5 MB (uncompressed)	МТ	JavaScript	Additional Information Version: 1.7
	Foun	2.0.1	<25 kB (core gzipped)	Apache 2 ^[3]	JavaScript	Updated: April 20, 2015
	any a	20 A00 2012				C) 4 071 110

4. In the popup window that appears, click the **Add extension** button.

web store			Add "Scrap	er"?		×	
	Extensions	It can:					
		Read	and change a	ll vour	data on the w	vebsites that you	
	S Data Mino	visit	2	1		,	
	A Data Will IC	Modif	y <mark>d</mark> ata that y	ou cop	y and paste		
Si offe	red by dvhtn	er Toois	10,920 USEIS		Cancel	Add extension	
OVERVIEW	REVIEWS		RELATED	þ			
1000/						Developer	🐑 Compatible
Comparison of Jav	aScript frameworks -	amework Version	Size	License	Source Language	☆ * ≡	
Wikipedia, the free	encyclopedia 🔨 1 Ar	pie 50K 0.9.3 1 Jul 2010	Variable, Core size: 40 k8 (minified & pripped)	MIT & CPL	2		Scraper gets da
T Selector	dive[1/10000/10/00]	pulad5 1.2.14 1 Mar 2014	36 k8 (minified & compressed)	MIT	JavaScript		and Into sprea

5. The extension is now added to Chrome.

Scraping Data from a Website

- 6. Now, open a new Chrome window.
- 7. Go to: <u>http://data.un.org/en/iso/tz.html</u>
- 8. Once the page opens, click on the at the **"Social Indicators"** section to show the table.
- 9. Now use the mouse pointer to highlight first three rows of the "Social Indicators" table, then right-click and select Scrape similar...(Note, if you select the entire table, the tool will not work. You have to select a portion of the table and Google will automatically find the rest of the table).

Social indicators			
	2005	2010	2017
Population growth rate ^{b,I} (average annual %)	2.8	3.1	3.1 ^c
Urban population ^b (% of total population)	Search 24. Print Scrape	DuckDuckGo for "2005 2010 2017 Population growth rateb,! (average" Ctrl+P	31.6 ^c
Urban population growth rate ^{b,I} (average annual %)	4.o	Ctri+Shift+	5.4 ^C
Fertility rate, total ^{b,1} (live births per woman)	5.7	5.6	5.2 ^c
Life expectancy at birth ^{b,I} (females/males, years)	55.4 / 52.0	60.1 / 57.5	64.8 / 60.8 ^C
Population age distribution ^b (0-14 and 60+ years, %)	45.3 / 4.6	45.3 / 4.7	44.9 / 4.7 ^a

10. In the Scraper window that opens, click the **Copy to clipboard** button.

Scraper	- data.	un.or	g/en/iso/tz.html			8
data.un.org/en/iso/tz.html			Column 1	Column 2	Column 3	Column 4
	_ /	1	Population growth rateb,I (average annual %)	2.8	3.1	3.1c
- Selector	1	2	Urban populationb (% of total population)	24.8	28.1	31.6c
XPath //details[4]/table/tbody/tr[td] ③ XPath Reference	1	3	Urban population growth rateb,I (average annual %)	4.8	5.4	5.4c
- Columns	1	4	Fertility rate, totalb,I (live births per woman)	5.7	5.6	5.2c
XPath Name	1	5	Life expectancy at birthb,I (females/males, years)	55.4 / 52.0	60.1 / 57.5	64.8 / 60.8c
ii *[1] Column 1 • •	1	6	Population age distributionb (0-14	45.3 / 4.6	45.3 / 4.7	44.9/
# *[2] Column 2	•		und oor years, voj			4.14
Presets Reset Scrape	•		Copy to clipt	board	kport to Goo	gle Docs

 Open up a new sheet in your spreadsheet software of choice such as Microsoft Excel or LibreOffice Calc and paste the copied content into cell A1.

Liberation Sans \bullet 10 \bullet a α a \underline{a} \underline{a} $\cdot \equiv \cdot \equiv \cdot \equiv = = $	% 0.0 📋	•00 <u>00</u> >= «	- E · F	· 🗖 · 📑							
A1 🔮 🛣 Σ = Column 1	-		1		1						
A Column 1	Column	B	Column	c 2	Column	1		E			
Column 1 Deputation growth rates I (overage enougl 0()	Column	2	Column	3		4					
Population growth rated, (average annual %)		2.8		3.1	3.10						
		24.8		28.1	31.60						
4 Urban population growth rateb, I (average annual %)		4.8		5.4	5.4c						
5 Fertility rate, totalb,I (live births per woman)		5.7		5.6	5.2c						
⁶ Life expectancy at <u>birthb</u> ,I (females/males, years)	55.4/	52.0	60.1/	57.5	64.8 /	60.8	IC				
7 Population age distributionb (0-14 and 60+ years, %)	45.3/	4.6	45.3/	4.7	44.9/	4.7	a				
⁸ International migrant stockm (000/% of total pop.)	770.8/	2.0	308.6/	0.7	261.2/	0.	5c				
⁹ Refugees and others of concern to UNHCR (000)	630.6n		273.8n		402.1e						
¹⁰ Infant mortality rateb, I (per 1 000 live births)		67.1		52.4	44.0c						
¹¹ Health: Total expenditure (% of GDP)		4.7		5.3	5.6d						
¹² Health: Physicians (per 1 000 pop.)	~0.00		~0.0p		~0.0q						
¹³ Education: Government expenditure (% of GDP)		4.6		4.6	3.5d						
¹⁴ Education: Primary gross enrol. ratio (f/m per 100 pop.)	101.0 /	106.2	99.3 /	98.6	82.9 /	80.5	iC				
¹⁵ Education: Secondary gross enrol. ratio (f/m per 100 pop.)	/		27.5/	34.8	30.8 /	33.7	'n				
¹⁶ Education: Tertiary gross enrol. ratio (f/m per 100 pop.)	0.9/	2.0h	1.9/	2.4	2.5/	4.9r					
¹⁷ Intentional homicide rate (per 100 000 pop.)				8.5	7.0c						
¹⁸ Seats held by women in national parliaments (%)		21.4		30.7		36.4					
19											
20											
21											
22											
23											

Export to Google Docs Option

You also have the option to export the HTML table to your Google Doc account. Here are the steps below.

12. In the Scraper window that opens, click the **Export to Google Docs...**

button.

uala.un.org/en/isonz.ntml				20 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0		Conulli
	1	1	Population growth rateb,I (average annual %)	2	3 3.1	4 3.1c
lector	1	2	Urban populationb (% of total population)	24.8	28.1	31.6c
⟨Path ▼ [//details[4]/table/tbody/tr[td] ② XPath Reference	1	3	Urban population growth rateb,l (average annual %)	4.8	5.4	5.4c
lumns	1	4	Fertility rate, totalb,I (live births per woman)	5.7	5.6	5.2c
XPath Name	1	5	Life expectancy at birthb,l (females/males, years)	55.4 / 52.0	60.1 / 57.5	64.8 / 60.8c
*[1] Column 1 • • • • • • • • • • • • • • • • • •		6	Population age distributionb (0-14 and 60+ years, %)	45.3 / 4.6	45.3 / 4.7	44.9 / 4.7a

13. If you are not signed in to Google, Chrome asks you to sing-in using your Gmail id. A window opens seeking permission for the Scraper extension. Click Allow.

Google		
	Control Contro	
	 Scraper would like to: 	
	View and manage Google Drive files and folders that you have opened or created with this app	1
	View and manage the files in your Google Drive	1
	By clicking Allow, you allow this app and Google to use your information in ac with their respective terms of service and privacy policies. You can change th other Account Permissions at any time.	cordance iis and
	Deny	Allow

14. A Google Spreadsheet now opens showing the Social Indicators data scraped from the UN website:

					1/4			A >	
5		100% - \$	% .0 .00 1	23 - Arial	-	10 -	B I S	<u>A</u> ↔. ⊞	23
fx	Column 1								
	А	В	С	D	E		F	G	
1	Column 1	Column 2	Column 3	Column 4					
2	Population growt	2.8	3.1	3.1c					
3	Urban population	24.8	28.1	31.6c					
4	Urban population	4.8	5.4	5.4c					
5	Fertility rate, tota	5.7	5.6	5.2c					
6	Life expectancy a	55.4 / 52.0	60.1 / 57.5	64.8 / 60.8c					
7	Population age d	45.3 / 4.6	45.3 / 4.7	44.9/ 4.7a					
8	International mig	770.8 / 2.0	308.6 / 0.7	261.2 / 0.5c					
9	Refugees and oth	630.6n	273.8n	402.1e					
10	Infant mortality ra	67.1	52.4	44.0c					
11	Health: Total exp	4.7	5.3	5.6d					
12	Health: Physiciar	~0.00	~0.0p	~0.0q					
13	Education: Gove	4.6	4.6	3.5d					
14	Education: Prima	101.0 / 106.2	99.3 / 98.6	82.9 / 80.5c					
15	Education: Secon	· /	27.5/ 34.8	30.8 / 33.7r					
16	Education: Tertia	0.9 / 2.0h	1.9/ 2.4	2.5/ 4.9r					
17	Intentional homic		8.5	7.0c					_
18	Seats held by wo	21.4	30.7	36.4					_
19									_
20									_
21									_
22									-
23									

Exercise: Scrape Data Using Browser Extensions

Use the Chrome browser to scrape data from the "List of Regions" table available on this Wikipedia page: <u>https://en.wikipedia.org/wiki/Regions_of_Tanzania</u>

Region	♦ Capital ♦	Districts •	Area (km²) 🕈	Population (2012) \$	Postcode \$	Zone 🔶	Мар
Arusha Region	Arusha	7	37,576	1,694,310	23xxx	Northern	
Dar es Salaam Region	Dar es Salaam	5	1,393	4,364,541	11xxx	Coastal	
Dodoma Region	Dodoma	7	41,311	2,083,588	41xxx	Central	
Geita Region	Geita	5	20,054	1,739,530	30xxx	Lake	
ringa Region	Iringa	5	35,503	941,238	51xxx	Southern Highlands	
Kagera Region	Bukoba	8	25,265	2,458,023	З5ххх	Lake	
Katavi Region	Mpanda	3	45,843	564,604	50xxx	Western	

Task 3: Scraping Data Using ImportHTML

The Google Spreadsheet function **=importHTML(URL, QUERY, INDEX)** helps scrape a table from an HTML web page into a Google spreadsheet. Within this function:

- URL: should be the target web page containing the table and MUST BE IN DOUBLE QUOTES.
- **QUERY:** should have the string "table" and **MUST BE IN DOUBLE QUOTES**.
- **INDEX:** should be a number the identifies the N'th table in the page (counting starts at 0).

As an example, let's use Google Sheets to download data about the regions of Tanzania.

- 1. Open <u>https://en.wikipedia.org/wiki/Regions_of_Tanzania</u> and scroll down the page to find the "**List of regions**" table.
- Now, open a blank Google Sheets from: <u>https://docs.google.com/spreadsheets/</u> (if you are not signed-in, Google will ask you to sign in).

Sheets						
	Start a new sproad	sheet				MORE C
		N10			The Street Trees	
	+		The second secon	1 =	The Prove Dense	
				<u>r-</u>		

3. In the new Google Sheet that opens, type the following function:
=importHTML("<u>https://en.wikipedia.org/wiki/Regions_of_Tanzania</u>", "table", 0)

	Untitled spreadsheet Image: Second secon															
5		5 7	100% -	£ %	.0 <u>,</u> .0 <u>0</u>	123 -		Ŧ	6	*	В	I	с <mark>р</mark>	Α		⊞
fx	=impor	tHTML("https://en.w	vikipedia	a.org/wi	ki/Reg	ions_of_Ta	anzania	", "tal	ole"	, 0)					
	A1 ,	4	В		С		D		E			F			G	
1 ?	=import	thtml("	https://en.w	ikipedia	.org/wik	i/Regi	ons_of_Tar	nzania'	', "tab	le",	0)					
2																
3																

4. Now press Enter and the Google Sheet populates data from the "List of regions' table from the Regions of Tanzania Wikipedia page.

5		100% - £	% .0, .00 1	23 - Arial	- 10 -	B <i>I</i> 5	<u>A</u> ♦. ⊞	53 · E · .	<u>↓</u> - ¦ - 🏷 -
fx	Arusha Region	1							
	А	В	С	D	E	F	G	н	I
1	Region	Capital	Districts	Area (km2)	Population (2012	Postcode	Zone	Мар	
2	Arusha Region	Arusha	7	37,576	1,694,310	23xxx	Northern		
3	Dar es Salaam F	Dar es Salaam	5	1,393	4,364,541	11xxx	Coastal		
4	Dodoma Region	Dodoma	7	41,311	2,083,588	41xxx	Central		
5	Geita Region	Geita	5	20,054	1,739,530	30xxx	Lake		
6	Iringa Region	Iringa	5	35,503	941,238	51xxx	Southern Highla	inds	
7	Kagera Region	Bukoba	8	25,265	2,458,023	35xxx	Lake		
8	Katavi Region	Mpanda	3	45,843	564,604	50xxx	Western		
9	Kigoma Region	Kigoma	8	37,040	2,127,930	47xxx	Western		
10	Kilimanjaro Regi	Moshi	7	13,250	1,640,087	25xxx	Northern		
11	Lindi Region	Lindi	6	66,040	864,652	65xxx	Coastal		
12	Manyara Region	Babati	6	44,522	1,425,131	27xxx	Northern		
13	Mara Region	Musoma	7	21,760	1,743,830	31xxx	Lake		
14	Mbeya Region	Mbeya	7	35,954	2,707,410[a]	53xxx	Southern Highla	inds	
15	Morogoro Region	Morogoro	7	70,624	2,218,492	67xxx	Coastal		
16	Mtwara Region	Mtwara	7	16,710	1,270,854	63xxx	Coastal		
17	Mwanza Region	Mwanza	7	9,467	2,772,509	33xxx	Lake		
18	Njombe Region	Njombe	6	21,347	702,097	59xxx	Southern Highla	inds	
19	Pemba North Re	Wete	2	574	211,732	75xxx	Zanzibar		
20	Pemba South Re	Chake Chake	2	332	195,116	74xxx	Zanzibar		
21	Pwani Region	Kibaha	7	32,547	1,098,668	61xxx	Coastal		
22	Rukwa Region	Sumbawanga	4	22,792	1,004,539	55xxx	Southern Highla	Inds	
23	Ruvuma Region	Songea	6	63,669	1,376,891	57xxx	Southern Highla	Inds	
24	Shinvanga Regio	Shinvanga	5	18.901	1.534.808	37xxx	Lake		