

Taller de manejo de datos con herramientas libres



Práctica Septima Clase

Scraping

Scraping es un término que, traducido al español, literalmente quiere decir “rascado”. Sin embargo, en este contexto, se refiere a la obtención o extracción de datos desde sitios web o archivos PDF para almacenarlos en formatos que sean más sencillo de utilizar posteriormente, como los CSV. Existen técnicas avanzadas de scraping que requieren programación, pero también existen muchas herramientas libres que, sin necesidad de programar, nos permiten extraer datos que sean de nuestro interés.

Caso Consumiendo tablas HTML desde Google Sheets.

Descargar los valores promedios anuales de los datos climáticos del aeropuerto de Santa Rosa desde la página Tutiempo.net

Una de las formas mas sencillas de obtener datos, es utilizando funciones de Google Sheets. La siguiente función permite extraer los datos de una tabla con formato HTML desde cualquier página web y la coloca dentro de una hoja de cálculo de Google.

=importHTML("","table",N)

La URL de la página web de destino y el elemento tabla a extraer deben estar entre comillas dobles: “”. El número N identifica la tabla N en la página web (comenzando desde 1) como la tabla destino para extraer los datos.

Así, por ejemplo, la función:

=importHTML("www.inta.gob.ar/anguil","table",2)

Se conectará con la página web www.inta.gob.ar/anguil, buscará el segundo elemento del tipo tabla (“table”) y lo cargará en la hoja de cálculo.

- 1 Ir al sitio web: <https://www.tutiempo.net/clima/ws-876230.html>, se puede ver que presenta datos climáticos de Santa Rosa:

Taller de manejo de datos con herramientas libres

The screenshot shows the TuTiempo.net website interface. At the top, there's a navigation bar with links for 'Clima', 'El tiempo' (selected), 'Astronomía', 'Más', and social media icons for Facebook, Twitter, and Google+. Below the navigation is a banner for 'Antideslizantes para pisos' (Non-slip mats for floors) from Johnson, featuring a person slipping on a wet floor. The main content area displays weather data for station 876230 (SAZR) in Santa Rosa, Argentina, with coordinates -36.56, -64.26 at 190m altitude. It includes a table of average and total annual climate values from 1958 to 1975.

Año	T	TM	Tm	PP	V	RA	SN	TS	FG	TN	GR
1958	-	-	-	-	-	-	-	-	-	-	-
1965	-	-	-	-	-	-	-	-	-	-	-
1966	-	-	-	-	-	-	-	-	-	-	-
1967	-	-	-	-	-	-	-	-	-	-	-
1968	15.9	23.0	8.6	-	11.6	65	1	36	27	0	0
1973	-	-	-	-	-	103	2	31	33	0	3
1974	15.6	22.7	8.0	538.24	13.0	110	2	53	27	0	6
1975	16.0	-	8.6	-	16.4	95	1	50	37	1	5

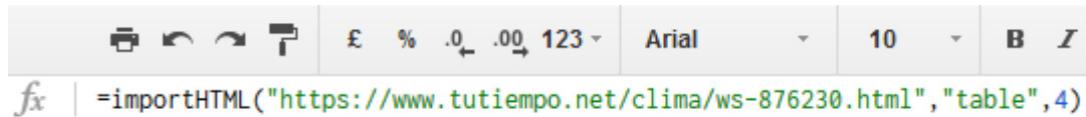
Ahora abrimos un nuevo documento de Google Sheets y le ponemos de nombre **Datos Climáticos Santa Rosa**.

The screenshot shows a Google Sheets document titled 'Datos climáticos Santa Rosa'. The window title bar includes tabs for 'CursoExtracurricular - ...', 'Datos climáticos Sant...', 'script-tmp-inta_barow_-_-...', and 'Top 30 F'. The URL in the address bar is https://docs.google.com/spreadsheets/d/1SEmcSOBLEXhLFFjs6CQ0T2CUw1QDVXmxB63UMDKrFrM/edit#. The spreadsheet has a green header row with column labels A through F. Row 1 contains data for the first column, and rows 2 and 3 are empty.

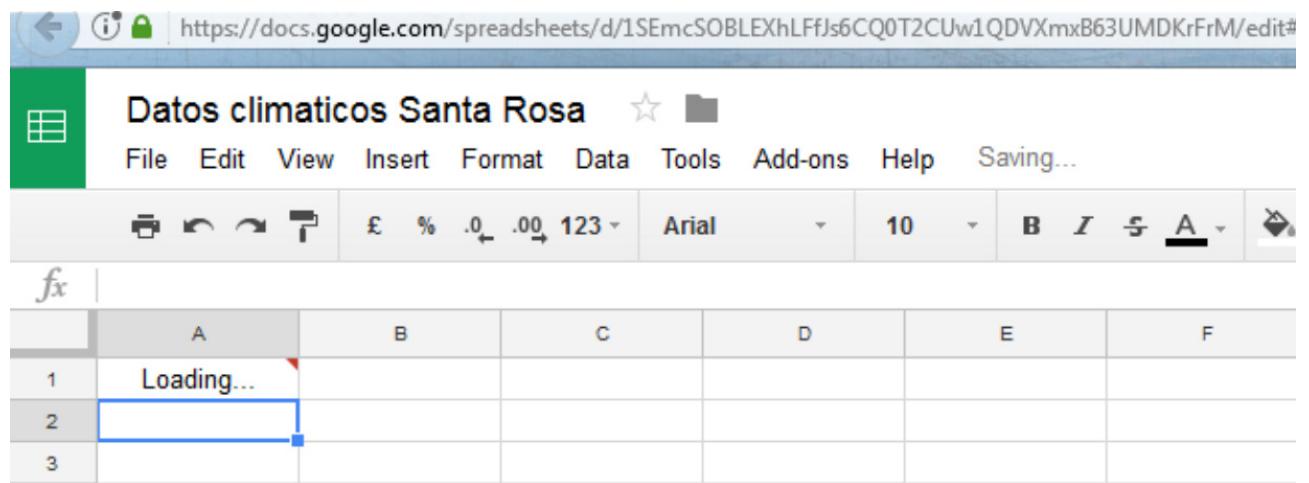
Taller de manejo de datos con herramientas libres

En la celda **A1**, colocamos la siguiente función y presionamos Enter:

```
=importHTML("https://www.tutiempo.net/clima/ws-876230.html","table",4)
```



Se presentará un cartel Cargando... (Loading...):



Y cuando finalice de traer los datos, presentará la tabla correspondiente:

	Año	T	TM	Tm	PP	V	RA	SN	TS	FG
1	*1958*	-	-	-	-	-	-	-	-	-
2	*1965*	-	-	-	-	-	-	-	-	-
3	*1966*	-	-	-	-	-	-	-	-	-
4	*1967*	-	-	-	-	-	-	-	-	-
5	*1968*	15.9	23	8.6	-	11.6	65	1	36	
6	*1973*	-	-	-	-	-	103	2	31	
7	*1974*	15.6	22.7	8	538.24	13	110	2	53	
8	*1975*	16	-	8.6	-	16.4	95	1	50	
9	*1976*	15.1	21.9	8	-	14.9	96	6	47	
10	*1977*	16.9	23.1	9.1	640.64	15.1	85	3	41	
11	*1978*	15.7	23	8.9	-	14.5	113	3	44	
12	*1979*	15.6	24	8.8	761.26	14.7	92	2	47	
13	*1980*	16.5	25.1	9.1	859.55	14	83	4	33	
14	*1981*	16.1	24.4	9.2	377.7	14.4	106	3	55	
15	*1982*	-	-	-	-	-	-	-	-	
16	*1983*	15.6	23.5	8.9	940.79	15.2	71	1	61	
17	*1984*	14.7	21.5	8.8	-	11.9	99	2	34	
18	*1985*	-	-	-	-	-	-	-	-	
19	*1986*	-	-	-	-	-	-	-	-	
20	*1987*	-	-	-	-	-	-	-	-	

Taller de manejo de datos con herramientas libres

¿Cómo sabemos qué número de tabla es la que queremos?, Probando!, iniciamos en 1 y vamos aumentando el número hasta que aparece la que nosotros necesitamos.

Para quedarnos con los datos de la tabla, sin necesidad de estar conectados, seleccionamos todos los datos, los copiamos (Ctrl+C) y luego seleccionamos **Edición -> Pegado Especial -> Pegar solo valores**

The screenshot shows a Google Sheets spreadsheet titled "Datos climaticos Santa Rosa". The "Edit" menu is open, and the "Paste special" option is selected. A submenu displays several options for pasting data:

- Paste values only (Ctrl+Shift+V)
- Paste format only (Ctrl+Alt+V)
- Paste all except borders
- Paste column widths only
- Paste formula only
- Paste data validation only
- Paste conditional formatting only
- Paste transpose

The main spreadsheet area shows a table with columns D, E, and F. The first column is labeled "Año" and contains years from 1958 to 1983. The second column is labeled "Tm" and has values like 15.6, 23.5, etc. The third column is labeled "PP" and has values like 8.9, 940.79, etc.

De esta manera, desaparece la función importHTML y queda la tabla con los valores descargados del sitio web.

- 2 Ahora debajo de los datos, traiga la tabla con las referencias y almacénela en la misma hoja. Debe quedar algo similar a esta pantalla:

Taller de manejo de datos con herramientas libres

48	*2014*	16.1	24.5	9.2	-
49	*2015*	16.2	24.5	9.2	906.7
50	*2016*	15.4	23.1	8.9	993.3
51	*2017*	-	-	-	-
52					
53	T	Temperatura media anual			
54	TM	Temperatura máxima media anual			
55	Tm	Temperatura mínima media anual			
56	PP	Precipitación total anual de lluvia y/o nieve derretida (mm)			
57	V	Velocidad media anual del viento (Km/h)			
58	RA	Total días con lluvia durante el año			
59	SN	Total días que nevó durante el año			
60	TS	Total días con tormenta durante el año			
61	FG	Total días con niebla durante el año			
62	TN	Total días con tornados o nubes de embudo durante el año			
63	GR	Total días con granizo durante el año			

En el artículo: <http://www.bigdatanews.com/profiles/blogs/top-30-free-web-scraping-software> se presentan más opciones de software para realizar Scraping. En la parte final de esta práctica se presenta el software Web Scraping.

Caso Obteniendo datos de un PDF

Existen varios software que realizan conversiones de archivos, algunos de ellos funcionan on-line lo que evita tener que instalar el sistema en la propia computadora. Dentro de este grupo tenemos PDFToExcel.

Transformar los datos de dos archivos PDF con tablas a un formato CSV legible.

- 1 Descargamos los dos archivos PDF que se encuentran en la carpeta **Prácticas -> Clase8** del disco virtual.

Para utilizar el aplicativo entramos en <https://www.pdftoexcelonline.com/>

La pantalla de inicio nos presenta un paso a paso, donde lo primero es seleccionar el archivo a convertir. Presionando el botón **Select your file** y seleccionamos el archivo **crns1701.pdf**, luego completamos con el mail al cual deseamos que nos envíe el archivo y finalmente presionamos el botón **Convert Now**.

Taller de manejo de datos con herramientas libres



1. Select your **PDF file** to convert

Select your file

2. Email converted file to:

your-email@example.com

Receive news, tips, and offers

3. Convert my PDF to Excel ?

Convert Now

By converting a file you agree to our [Terms of Service](#).

1. Select your **PDF file** to convert

✓ crns1701.pdf

Select another file »

2. Email converted file to:

yabellini@gmail.com

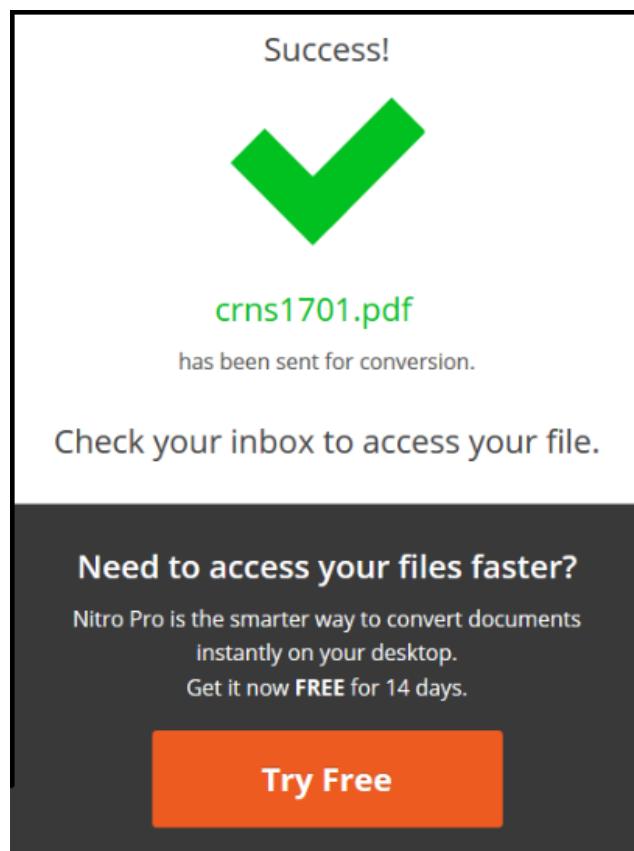
Receive news, tips, and offers

3. Convert my PDF to Excel ?

Convert Now

By converting a file you agree to our [Terms of Service](#).

Nos aparece un cartel como el siguiente donde nos indica que nos enviarán el archivo convertido por mail:



Taller de manejo de datos con herramientas libres

Entrando al mail que indicamos encontraremos un correo con un link para descargar el archivo convertido, lo guardamos y abrimos para analizar la transformación.

The screenshot shows a Firefox browser window. At the top, there's a notification from Nitro Cloud: "Here's your converted document - Nitro Here's your converted file! crns1701 converted successfully". Below it, a download dialog box is open, titled "Abriendo crns1701.xlsx". The dialog shows the file has been chosen to open with Microsoft Excel (selected) and offers to save it instead. A checkbox for automatic future opening is also present. At the bottom are "Aceptar" and "Cancelar" buttons. A large orange button labeled "Download the document" is overlaid at the bottom of the browser window.

También existen software que se instalan en la computadora, uno de ellos es **Tabula** (<http://tabula.technology/>). Para utilizarlo descargamos el archivo comprimido que se encuentra en la carpeta software del disco virtual **tabula-win-1.1.1.zip**, lo copiamos en el escritorio, lo descomprimimos y luego vamos a la carpeta que acaba de extraer. Dentro de la misma se encuentran una serie de archivos, ejecute (haciendo doble click o presionando Enter) el programa "Tabula" que se encuentra dentro. Se abrirá un navegador web. Si no lo hace, abra su navegador web y vaya a **http://localhost: 8080**. Aparecerá una pantalla similar a la siguiente.

Taller de manejo de datos con herramientas libres



The screenshot shows the Tabula application running on a local server at localhost:8080. The main heading is "Import one or more PDFs". Below it is a "Browse..." button and an "Import" button. A welcome message "First time using Tabula? Welcome!" is displayed, followed by a "How to Use Tabula" section with five steps. Step 1: Upload a PDF file containing a data table. Step 2: Select the table by clicking the top left corner of a table and dragging the mouse to the bottom right corner, until all of the data is included in the shaded selection area. Step 3: A window will then appear containing your data. Inspect the data to make sure it looks correct. If data is missing, you may have to slightly expand your selection. Step 4: Click the Download button. Step 5: Now you can work with your data as text file or a spreadsheet rather than a PDF! (You can open the downloaded file in Microsoft Excel or the free LibreOffice Calc).

Note: Tabula only works on text-based PDFs, not scanned documents.

Having trouble with Tabula?

1. Tabula said "Sorry, your PDF file is image-based" -- what does that mean? Your PDF does not have any embedded text. It might have been scanned from paper. Tabula is not able to extract any data from image-based PDFs. You can try OCRing the PDF

Para iniciar la conversión, presionamos el botón **Browse**, luego elegimos el archivo **santa2016-12.pdf** y presionamos el botón **Import**. Inicia la conversión del archivo PDF.

The screenshot shows the Tabula application running on a local server at localhost:8080. The main heading is "Import one or more PDFs". Below it is a "Browse..." button with the file path "santa2016-12.pdf" and an "Import" button. A progress bar indicates the upload status: "santa2016-12.pdf waiting to be processed..."

Cuando termina nos presenta una pantalla para trabajar, si presionamos el botón **Autodetect Tables** nos marca la sección a transformar y con el botón **Preview & Export Extracted Data** podemos obtener la información seleccionada.

Taller de manejo de datos con herramientas libres

The screenshot shows the Tabula software interface. At the top, there's a navigation bar with links for 'Tabula', 'My Files', 'About', 'Help', and 'Source Code'. Below the navigation bar, a file named 'santa2016-12.pdf' is open. A red dashed box highlights a specific table in the document. On the right side, a preview window displays the extracted data from the selected table. The preview window has a header row with columns labeled 'Día', 'T', 'TM', 'Tm', 'SLP', 'H', and 'PP'. The data rows show various values for each column. Below the main table, there are two additional sections: 'Medias y totales mensuales' and another set of data rows.

Día	T	TM	Tm	SLP	H	PP
24	26.6	34.4	16.4	1006.4	42	0
8	17.7	25.5	10	1014.9	67	0
1	25.6	34.6	16.8	1009.2	49	0
14	14	22.5	10.2	1013.8	80	4.06
1	12.6	20.2	7.3	1021.8	84	5.08
3	20.1	30.8	10	1012.1	40	0
18	21.8	35.9	14.4	1010.2	55	13.97
29	23.1	30.6	13.2	1019.5	45	0
22	15.6	24.1	6.8	1015.1	74	0
15	25.4	36.3	16	1012.2	36	0
26	20.6	30.9	16.5	1007.8	72	14.99
1	14.2	26.5	10	1008.2	55	8.89
4	14.4	24.8	12	1016.3	87	39.88
23	19.1	24.3	13.9	1014.3	74	0
Medias y totales mensuales						
15	16.1	22.4	10.8	1010.2	85	40.89
19	20.3	29.7	11.9	1009.8	54	0
27	17.6	32.6	6.7	1015.5	85	10.00

La pantalla siguiente nos presenta los datos y nos da varias opciones para exportarlos o copiarlos al portapapeles. También podemos volver atrás para cambiar la selección si vemos que lo elegido no corresponde con lo que necesitamos extraer.

Lo exportaremos a un archivo CSV, seleccionamos ese formato en **Export Format** y presionando en el botón **Export**.

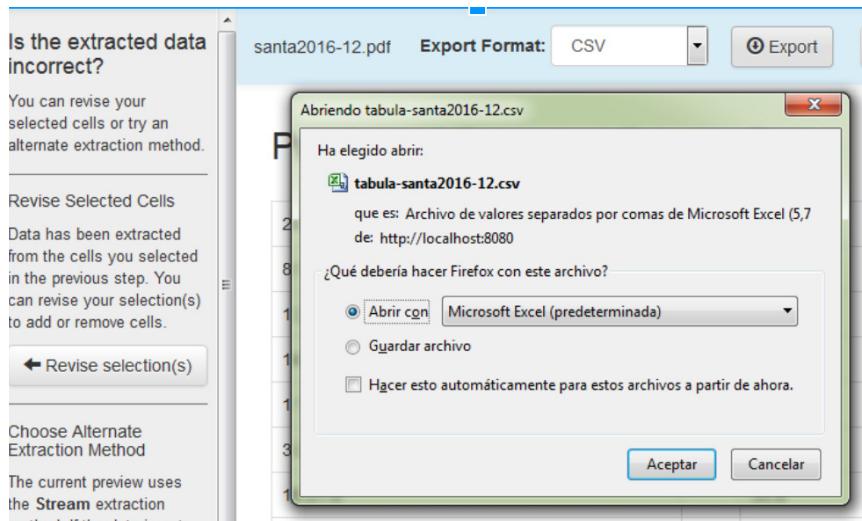
The screenshot shows the Tabula software interface with the 'Export Format' dropdown set to 'CSV'. A preview window titled 'Preview of Extracted Tabular Data' displays the same data as the previous screenshot. To the left, there's a sidebar with several options:

- 'Is the extracted data incorrect?' with a note about revising selected cells or trying an alternate extraction method.
- 'Revise Selected Cells' button.
- 'Choose Alternate Extraction Method' section with notes about Stream and Lattice methods.
- 'Stream' and 'Lattice' buttons.

	24 26.6	34.4	16.4	1006.4	42	0
8	17.7	25.5	10	1014.9	67	0
1	25.6	34.6	16.8	1009.2	49	0
14	14	22.5	10.2	1013.8	80	4.06
1	12.6	20.2	7.3	1021.8	84	5.08
3	20.1	30.8	10	1012.1	40	0
18	21.8	35.9	14.4	1010.2	55	13.97
29	23.1	30.6	13.2	1019.5	45	0
22	15.6	24.1	6.8	1015.1	74	0
15	25.4	36.3	16	1012.2	36	0

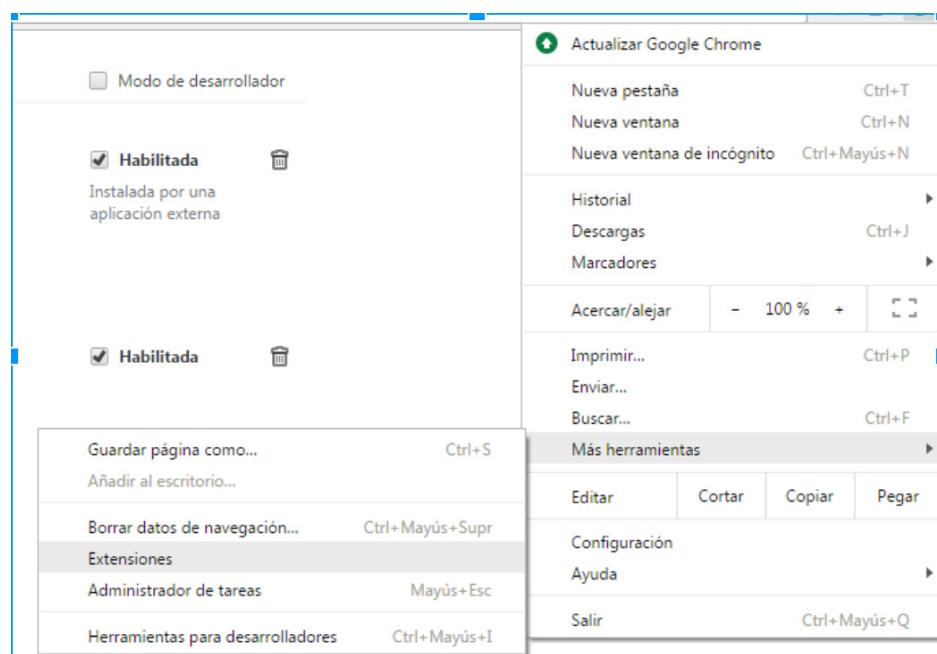
Taller de manejo de datos con herramientas libres

Si seleccionamos Guardar y damos Aceptar el archivo se almacenará en nuestro disco. Lo guardamos y abrimos para analizar la transformación.



Bonus: Web Scrapper

Es una extensión gratuita para Google Chrome, para instalarla Presionamos en **Herramientas -> Más herramientas -> Extensiones** y hacemos click en **Obtener más extensiones**.



En la tienda de extensiones de Google Chrome, buscamos **Web Scraper** y seleccionamos + AÑADIR A CHROME si la extensión ya se encuentra instalada, se presenta el cartelito de añadido:



Taller de manejo de datos con herramientas libres



The screenshot shows the Chrome Web Store interface. In the search bar at the top left, 'web scraper' is typed. On the right side of the header, there are buttons for 'Iniciar sesión' and a gear icon. Below the search bar, the word 'Extensiones' is displayed. The main content area shows two extension cards:

- Web Scraper** by Martins Balodis: A tool for data extraction from websites. It has a green 'VALORAR' button, a 'Productividad' rating, and a 4-star review count of 312.
- Advanced Web Scraper** offered by https://www.datascraping.co: An easy, powerful web scraping app for screen scraping using CSS. It has a blue '+ AÑADIR A CHROME' button, a 'Herramientas para desarrolladores' rating, and a 5-star review count of 101.

Para utilizar Web Scraper, primero navegamos hacia la página que queremos obtener los datos, en este caso <https://www.tutiempo.net/clima/ws-876230.html>.

Para realizar scrapping siempre es conveniente analizar como el sitio web genera las páginas que nos presenta, si vemos tutiempo.net presenta un listado de años con datos disponibles, si hacemos click en uno de esos años, por ejemplo, **1977**, vemos que la url cambió, agregando el año antes del código de estación, a su vez si seleccionamos un mes, por ejemplo, **Noviembre**, vemos que la url cambia nuevamente, agregando el mes al año:

Url original: <https://www.tutiempo.net/clima/ws-876230.html>

Url de un año específico: <https://www.tutiempo.net/clima/1977/ws-876230.html>

Url de un mes y un año específico: <https://www.tutiempo.net/clima/11-1977/ws-876230.html>

Con esta información es más sencillo poder obtener los datos, ya que conocemos la estructura de cómo se presentan los mismos en la web. Para iniciar Web Scraper se presiona **F12** o bien **Herramientas (Tools) -> Herramientas para Desarrolladores (Developer Tools)** y se busca Web Scraper en los tabs superiores:

Taller de manejo de datos con herramientas libres

The screenshot shows a Chromium browser window with the following interface elements:

- Toolbar:** Includes icons for extensions, history, and tabs.
- Address Bar:** chrome://extensions/?id=lmebglmnljbaodajnmnopkakhklgcail
- Extensions Panel:** Shows "Web Scraper 0.1.4.7" as installed, with options to allow incognito mode or file URLs.
- Help Panel:** Shows "Get more extensions".
- Right-click Context Menu:** Opened from the extensions panel, with the following items:
 - New tab (Ctrl+T)
 - New window (Ctrl+N)
 - New incognito window (Ctrl+Shift+N)
 - Bookmarks
 - Recent Tabs
 - Edit (Cut, Copy, Paste)
 - Zoom (- 100% +)
 - Save page as... (Ctrl+S)
 - Find... (Ctrl+F)
 - Print... (Ctrl+P)
 - Tools** (highlighted in orange)
 - History (Ctrl+H)
 - Downloads (Ctrl+J)
 - Sign in to Chromium...
 - Settings
 - About Chromium
 - Help
 - Exit (Ctrl+Shift+Q)
- Developer Tools:** A modal window titled "Create application shortcuts..." with a list of developer tools:
 - Extensions
 - Task manager
 - Clear browsing data... (Shift+Esc)
 - Encoding
 - View source (Ctrl+U)
 - Developer tools** (highlighted with a red box and number 3)
 - JavaScript console (Ctrl+Shift+J)
 - Inspect devices
- Sitemaps Tab:** Shows a table of sitemaps with columns for ID, Start URL, and actions (Browse, Delete). Two entries are listed: "ecommerce" (Start URL: http://example-commerce.com/) and "news" (Start URL: http://example-news.com/).
- Bottom Navigation:** Elements, Network, Sources, Timeline, Profiles, Resources, Audits, Console, Web Scraper (highlighted with a red box and number 4), Sitemaps, Sitemap (dropdown), Create new sitemap (dropdown).

Para iniciar el scraping hacemos clik en **Create new sitemap -> Create sitemap**

The screenshot shows the developer tools Sitemaps tab with the following interface elements:

- Table:** Shows a grid of data with columns labeled 1980, 1981, 16.0, 20.1, and 9.0.
- Bottom Navigation:** Elements, Console, Sources, Network, Performance, Sitemaps, Sitemap (dropdown), Create new sitemap (dropdown).
- Create Sitemap Dialog:** A modal window with two buttons:
 - Create sitemap** (highlighted with a blue background)
 - Import sitemap

Taller de manejo de datos con herramientas libres



Le ponemos de nombre **tutiemposantarosa2016**. Para generar la url de inicio, vamos a aprovechar la numeración que realiza tutiempo.net, nuestro objetivo es descargar los datos diarios de Diciembre del 2016, por lo que la url quedaría conformada de la siguiente manera: <https://www.tutiempo.net/clima/12-2016/ws-876230.html>

Si quisieramos descargar los datos de Enero a Diciembre del 2016, la url quedaría conformada de la siguiente forma: [https://www.tutiempo.net/clima/\[01-12\]-2016/ws-876230.html](https://www.tutiempo.net/clima/[01-12]-2016/ws-876230.html)

Sitemap name

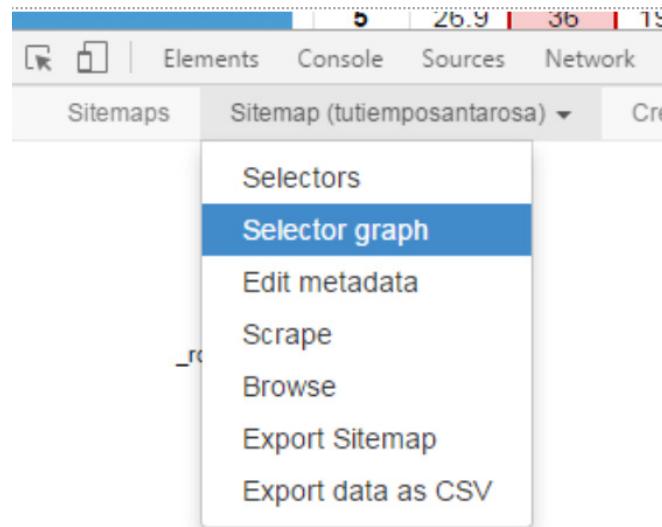
Start URL

Ahora que tenemos generado la url inicial, empezamos a configurar los elementos, para ello creamos un nuevo selector, del tipo **tabla**, ponemos de nombre DatosDiarios, Presionamos **Select** y hacemos click en la tabla y presionamos el botón **Done Selecting!**, chequeamos que se haya seleccionado correctamente los títulos y las columnas, seleccionamos **Multiple** y en la sección de columnas (**table column**) destildamos las columnas VG, RA, SN, TS y FG. Finalmente, guardamos el selector, haciendo click sobre el botón **Save Selector**.

The screenshot shows the configuration for a new selector named "Datos Diarios". The "Type" is set to "Table". The "Selector" field contains "table.medias". The "Header row selector" is set to "tr:nth-of-type(1)". The "Data rows selector" is set to "tr:nth-of-type(n+2)". The "Multiple" checkbox is checked. The "Delay (ms)" field is empty. The "Parent Selectors" field contains "_root". In the "table columns" section, four columns are listed: "Día", "T", "TM", and "Tm". Each column has a "Result key" field: "Día", "T", "TM", and "Tm". All four checkboxes under "Include into result" are checked.

Taller de manejo de datos con herramientas libres

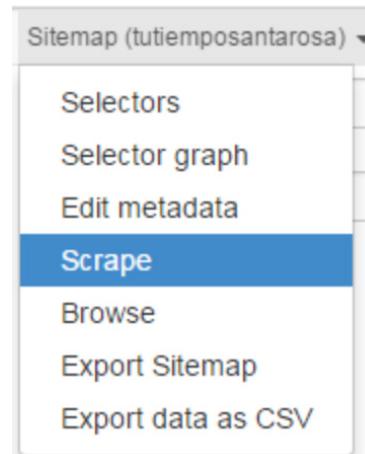
Podemos analizar cómo va quedando configurado nuestro mapa de scraping seleccionando Sitemap -> Selector graph



Nos presenta un esquema gráfico de nuestro scrapper:



Para iniciar la captura de los datos seleccionamos:

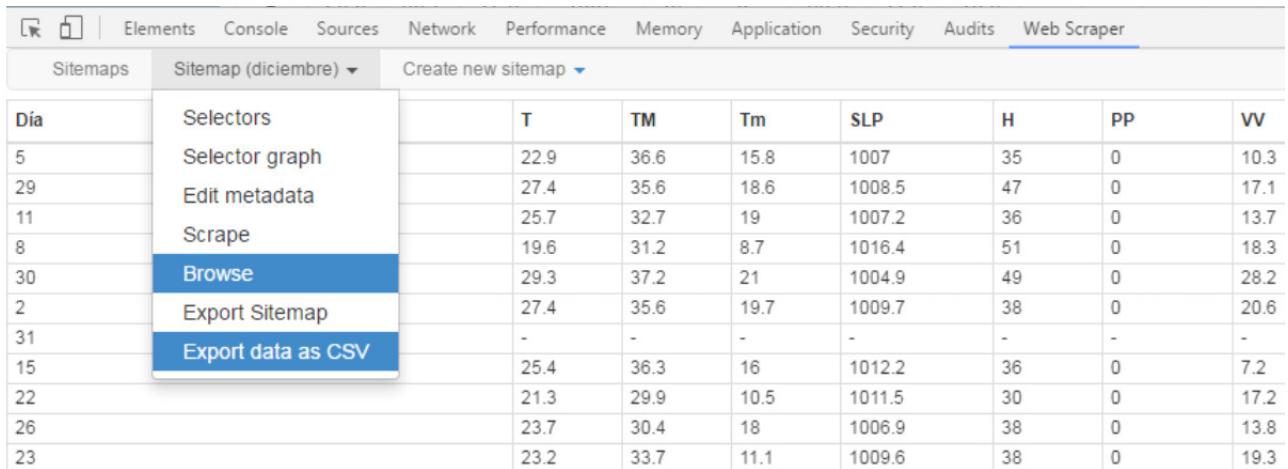


Dejamos los valores por defecto y presionamos **Start Scrapping**

A screenshot of the Chrome DevTools interface showing the Sitemap panel. A configuration dialog is open with the following settings: Request interval (ms) set to 2000, Page load delay (ms) set to 500, and a large blue 'Start scraping' button at the bottom.

Taller de manejo de datos con herramientas libres

Aparece una pantalla que simula la navegación y va realizando la captura de los datos correspondientes al mapa que armamos. Cuando finaliza nos presenta unas tablas con los datos descargados y nos permite almacenarlo en formato CSV seleccionando **Sitemap -> Export data as CSV**



Día	Selectors	T	TM	Tm	SLP	H	PP	VV
5	Selector graph	22.9	36.6	15.8	1007	35	0	10.3
29	Edit metadata	27.4	35.6	18.6	1008.5	47	0	17.1
11	Scrape	25.7	32.7	19	1007.2	36	0	13.7
8	Browse	19.6	31.2	8.7	1016.4	51	0	18.3
30	Export Sitemap	29.3	37.2	21	1004.9	49	0	28.2
2	Export data as CSV	27.4	35.6	19.7	1009.7	38	0	20.6
31		-	-	-	-	-	-	-
15		25.4	36.3	16	1012.2	36	0	7.2
22		21.3	29.9	10.5	1011.5	30	0	17.2
26		23.7	30.4	18	1006.9	38	0	13.8
23		23.2	33.7	11.1	1009.6	38	0	19.3

Cuando tiene el archivo preparado para descarga nos presenta un link **Download now!**, al hacer click nos descarga el archivo generado.

Export santa2016-12 data as CSV.

Waiting for the download button to appear. > [Download now!](#)

Para ver más detalles de cómo utilizar web scrapper se puede ver su documentación: <http://webscraper.io/documentation#installation>.