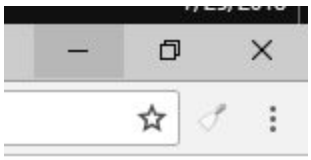
Web Scraping with the Chrome Scraper Extension

Set up the Scraper Extension

Install the Scraper extension in Chrome.

Test the Scraper extension

- [Highlight text and] **Right click** anywhere an choose Scrape similar…
- Highlight text and click the **icon in the toolbar** and choose Scrape similar... Instructions use right click, but you can always click the icon in the toolbar.



Check that the settings are correct

- Examine the window that opens, looking for **XPath** in the upper right. If it says “JQuery”, click the click the Reset button at the bottom of the Scraper window and then restart Chrome.

Using the Developer Toolbar in Chrome

- To Inspect an element, it right click on it.
- You can also press Command+Shift+C (Mac) or Ctrl+Shift+C (PC) then find the element.
- Right click on a tag and follow the menus to copy the XPath.
- In the Developer Window you can press Ctrl-F (PC) or Command+F (Mac) to open the Find toolbar and test your XPaths.

XPath Reference

HTML	**XPath** to locate **this**
this	p**/**span
anything	p**/**@class
this	p**[**@class**]**
this	p**[**@class="me"**]**
<p class=”me you”>**this**	p**[**contains**(**@class, "me"**)]**
this	p**[**not**(**@class="me"**)]**
me	p**[**text**()**="me"**]/**@class
this	*[@id=”me”]
Word**this**	div/text()
<hr>Word **this**	hr/following-sibling::p
<dl><dt>Word</dt><dd>**this**</dd></dl>	dl/dt[text()="Word"]/../dd
this	div//b

Adapted by GWU Libraries from: Wendy Mann, George Mason University Libraries, 2018.

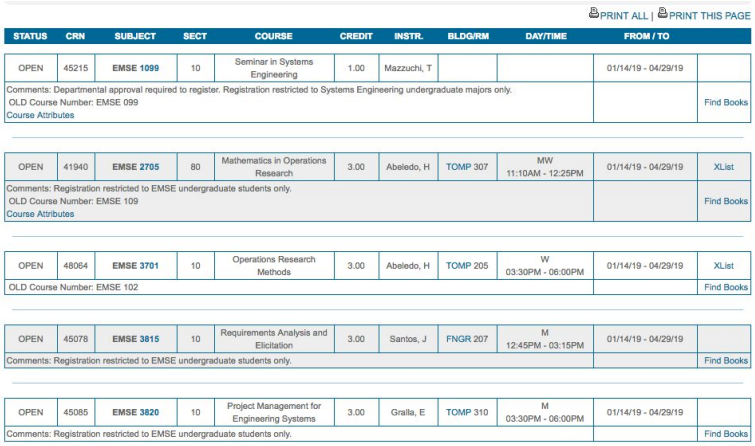
Scrape a List

Obtain the List of Courses and URLs

1. Go to https://my.gwu.edu/mod/pws/
2. In the Spring 2019 box, click Main Campus. Click on any department name.
3. **Highlight** a row and **right-click.**
4. Choose **Scrape similar...**
5. In the Scraper window, make sure XPath is selected on the left side. Adjust the XPath to get the data you need. For example:

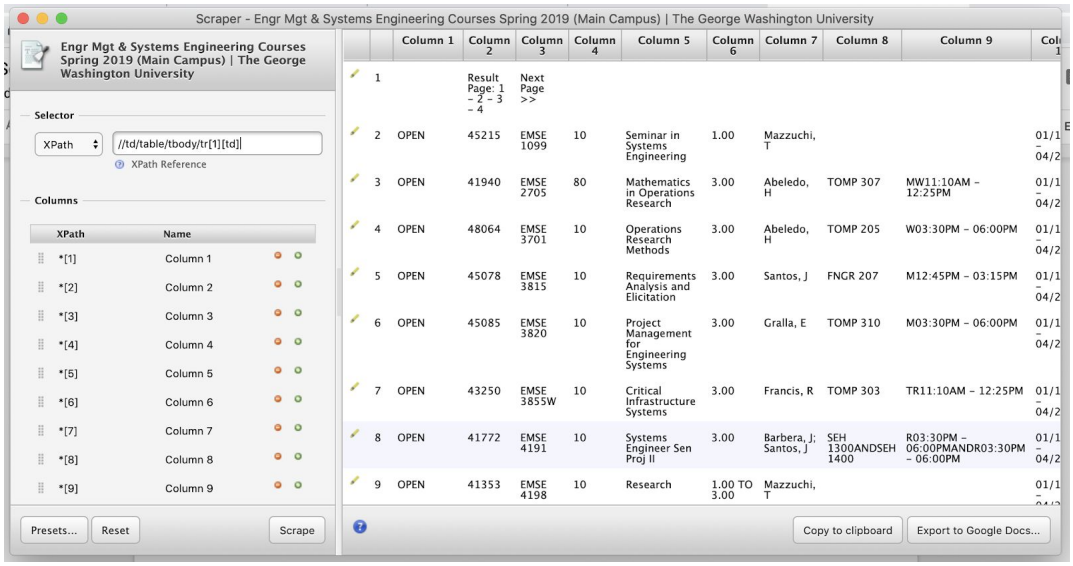
   ```
   //td/table[2]/tbody/tr[td]
   //table[@class="courseListing"]/tbody/tr[td]
   ```

6. Then press **Export to Google Sheets** or **Copy to clipboard** to paste to Excel.
7. In Excel or Google Sheets, choose to **Paste** (ex Ctrl-V)



PRINT ALL | PRINT THIS PAGE

STATUS	CRN	SUBJECT	SECT	COURSE	CREDIT	INSTR.	BLDG/RM	DAY/TIME	FROM / TO	
OPEN	45215	EMSE 1099	10	Seminar in Systems Engineering	1.00	Mazzuchi, T			01/14/19 - 04/29/19	
Comments: Departmental approval required to register. Registration restricted to Systems Engineering undergraduate majors only. OLD Course Number: EMSE 099 Course Attributes										Find Books
OPEN	41940	EMSE 2705	80	Mathematics in Operations Research	3.00	Abeledo, H	TOMP 307	MW 11:10AM - 12:25PM	01/14/19 - 04/29/19	XList
Comments: Registration restricted to EMSE undergraduate students only. OLD Course Number: EMSE 109 Course Attributes										Find Books
OPEN	48064	EMSE 3701	10	Operations Research Methods	3.00	Abeledo, H	TOMP 205	W 03:30PM - 06:00PM	01/14/19 - 04/29/19	XList
OLD Course Number: EMSE 102										Find Books
OPEN	45078	EMSE 3815	10	Requirements Analysis and Elicitation	3.00	Santos, J	FNGR 207	M 12:45PM - 03:15PM	01/14/19 - 04/29/19	
Comments: Registration restricted to EMSE undergraduate students only.										Find Books
OPEN	45085	EMSE 3820	10	Project Management for Engineering Systems	3.00	Gralla, E	TOMP 310	M 03:30PM - 06:00PM	01/14/19 - 04/29/19	
Comments: Registration restricted to EMSE undergraduate students only.										Find Books



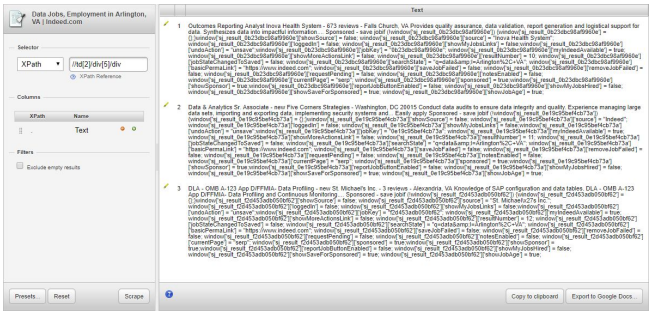
	Column 1	Column 2	Column 3	Column 4	Column 5	Column 6	Column 7	Column 8	Column 9	Col 1
1		Result Page: 1 - 2 - 3 - 4	Next Page >>							
2	OPEN	45215	EMSE 1099	10	Seminar in Systems Engineering	1.00	Mazzuchi, T			01/1 - 04/2
3	OPEN	41940	EMSE 2705	80	Mathematics in Operations Research	3.00	Abeledo, H	TOMP 307	MW11:10AM - 12:25PM	01/1 - 04/2
4	OPEN	48064	EMSE 3701	10	Operations Research Methods	3.00	Abeledo, H	TOMP 205	W03:30PM - 06:00PM	01/1 - 04/2
5	OPEN	45078	EMSE 3815	10	Requirements Analysis and Elicitation	3.00	Santos, J	FNGR 207	M12:45PM - 03:15PM	01/1 - 04/2
6	OPEN	45085	EMSE 3820	10	Project Management for Engineering Systems	3.00	Gralla, E	TOMP 310	M03:30PM - 06:00PM	01/1 - 04/2
7	OPEN	43250	EMSE 3855W	10	Critical Infrastructure Systems	3.00	Francis, R	TOMP 303	TR11:10AM - 12:25PM	01/1 - 04/2
8	OPEN	41772	EMSE 4191	10	Systems Engineer Sen Proj II	3.00	Barbera, J; Santos, J	SEH 1300ANDSEH 1400	R03:30PM - 06:00PMANDR03:30PM - 06:00PM	01/1 - 04/2
9	OPEN	41353	EMSE 4198	10	Research	1.00 TO 3.00	Mazzuchi, T			01/1 -

Adapted by GWU Libraries from: Wendy Mann, George Mason University Libraries, 2018.

Scrape Jobs from Indeed.com

1. **Go to https://www.indeed.com/** and search for jobs (I used “data”).

2. With your mouse, select the *whole* job listing, then right click and pick **Scrape Similar…**



   a. There should be a row for each job as in the screenshot. Ignore for now that only a few jobs appear.

3. **Identify an XPath** to obtain the **Job Title**

   a. To get you started, here is sample HTML for a title with some elements that you can use highlighted, see if you can use each one (e.g., make 4 different XPaths that work). Remember, classes are not separated, so use the contains() function to reference one.

```
<a target="_blank" id="sja1" data-tn-element="jobTitle"
class="jobtitle turnstileLink"
href="https://www.indeed.com/viewjob?jk=c74e528069aba
&amp;from=tp-serp&amp;tk=1c3gl72080j0f119&amp;tk=1c3gl0j0f119&amp;
jsa=6104" title="Digital Network Analyst" rel="noopener nofollow"
onmousedown="sjomd('sja1'); clk('sja1');"
onclick="setRefineByCookie([]);sjoc('sja1',0);
convCtr('SJ')">Digital Network Analyst</a>
```

4. **Add more columns** with the Green + button and identify XPaths to locate the other job elements, such as URL, Company, Location, and Summary

5. **Fix the XPath selector** to include all the jobs on the page by Inspecting the job listing as a whole and finding a more specific identifier for each individual job listing.

   a. The issue is that the Sponsored jobs are all in a special, extra div.

   b. **Note**: You may also have to fix some XPaths that previously worked because the structure of sponsored and unsponsored jobs also differs (uses div instead of span, etc)

6. **Display more job listings** by opening the Advanced Job Search (next to the Find button). Scroll to the bottom and have it display 50 results, find jobs and re-scrape.

7. **Remove extra spaces** in the Company and Summary columns
(to see why, feel free to copy the data to Excel first)

   a. The XPath function to do this is called normalize-space(). Put the path inside the ().

   b. You could also use the TRIM function in Excel or Find-and-Replace

“Answers” - Getting Jobs from Indeed.com

3. Any of these will work:

`.//a`	Reference an element
`.//@title`	Reference an attribute
`.//*[@data-tn-element='jobTitle']`	Filters to tags with a specific attribute value
`.//*[contains(@class,"jobtitle")]`	Filters to tags with part of an attribute value

4. Here are some examples

```
.//@href
.//*[@class="company"]
.//*[@class="location"]
.//*[@class="summary"]
```

5. Because the class for each job listing is “row result clickcard”, you must use the contains function:

```
//div[contains(@class,'clickcard')]
```

7a. Remove extra spaces with

```
normalize-space(.//*[@class="summary"])
```

7b. In Excel

```
=TRIM(C2)
```
Useful Links

Selectors

XPath -- https://www.w3schools.com/xml/xpath_syntax.asp and https://www.w3.org/TR/xpath/all/
CSS -https://www.w3schools.com/cssref/css_selectors.asp and http://learnlayout.com/display.html
JQuery - https://www.w3schools.com/jquery/jquery_ref_selectors.asp
Regular Expressions - http://infoguides.gmu.edu/data-work/regex

JQuery vs XPath: https://www.ibm.com/developerworks/library/x-xpathjquery/index.html

https://genius.com/Mat-brown-xpath-is-actually-pretty-useful-once-it-stops-being-confusing-annotated
CSS vs XPath: https://en.wikibooks.org/wiki/XPath/CSS_Equivalents

XPath Hierarchies

http://dh.obdurodon.org/introduction-xpath.xhtml
http://dh.newtfire.org/explainXPath.html
https://dpastov.blogspot.com/2015/10/preceding-sibling-and-following-signling-xpath.html

Reference

Tester: https://extendsclass.com/xpath-tester.html or http://xpather.com

Getting Started with Tabula

This exercise is an introduction to data scraping by using Tabula to scrape tables from a PDF document.

For this exercise, it is assumed that you know:

- the structure of a data file
    - each variable is a column
    - each observation is a row
- how to open or import a text delimited file into your spreadsheet software of choice
- what delimited file formats are (CSV, TSV, etc.)
- how to unzip or uncompress a file.

If you are unfamiliar with the above -- please ask for help!

Download and install Tabula

a.) http://tabula.technology
b.) Download the version appropriate for your operating system.
c.) Note: Windows users will need a copy of Java installed. https://www.java.com/download/

Tabula runs through your browser. If Tabula does not launch, try: **http://localhost:8080** or **127.0.0.1:8080**

Upload a PDF

Browse to and import PDF file containing a data table. For PDFs from websites, you need to download it and then open it in Tabula.

Select Tables

Browse to the page you want, then select the table(s) by clicking and dragging to draw a box around the table. Repeat selections to capture on subsequent pages.

Preview and Extract Tables:

Click "**Preview & Export Extracted Data**." Tabula will extract the data and display a preview.

- **Inspect the data to make sure it looks correct.** If data is missing or not placed in the appropriate cells, you can go back to adjust your selection.

In left margin see options to:

Revise selection

Change extraction method from **Stream** to **Lattice**.

Export

Once everything looks good, click the **Export** button.

Open Exported File

After saving the exported spreadsheet file, you can open the downloaded file in Microsoft Excel, Numbers, the free LibreOffice Calc, etc.)

Review Exported File

You may need to do some clean up on your spreadsheet. Tabula does not always grab the data perfectly.

Advanced Features

You can create a **Template** if working with multiple PDFs that have the same layout.