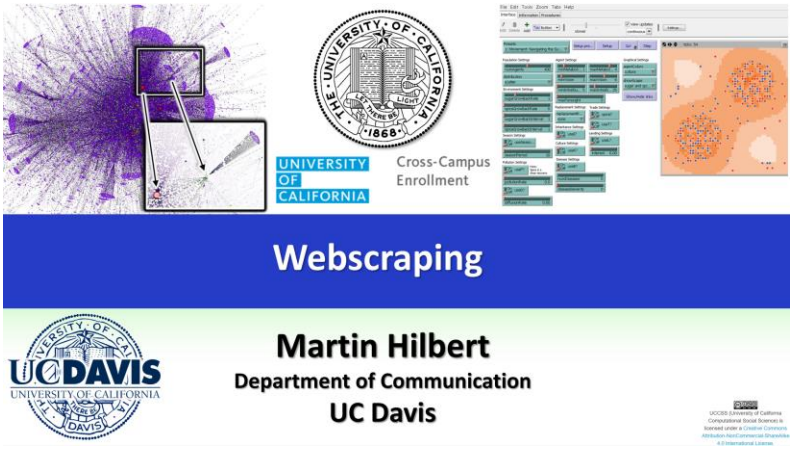


YouTube Titles

Contents

1. Installation

We use a Web Scraper that takes advantage of the Chrome browser from Google (it is a Chrome extension). Therefore, you need to install the Chrome browser and its extension. Please download and install:

1. Chrome Browser: https://www.google.com/chrome/index.html
2. From within Chrome, download and install: Web Scraper Chrome extension: http://webscraper.io/

Note: Top right corner of the Chrome window should now show: 

2. Building the Scraper

We now construct a scraper for the website we want to scrape. As a general rule, if you want to be able to scrape it, you need to be able to "copy-paste" it by hand as well. The other way around, it is very useful to first go to the site and by hand copy-paste the information you want, so you understand what you want and if you can get it or not.

In this case, we are interested in what kind of videos are featured on a certain YouTube channel and what kind of videos are recommended to people who watch the channel. More concretely, we want to scrape two types of content in a consecutive manner:

a. the **featured videos** on a YouTube channel

b. the **recommended videos** within each of these featured videos

At the end, we want to have a database that features the following variables:

	Featured videos		Recommended videos	
YouTube Channel	Name of video	URL of video	Name of video	URL of video

a. Featured Videos

Open up the website you are interested in to scrap via Chrome. We will use videos featured by the *The Young Turks* YouTube channel as an example for this guide (at the time, the most popular YouTube news channel in the US).

https://www.youtube.com/user/TheYoungTurks/videos

We want to get the **title** and **URL** of videos on this page (see Figure 1).

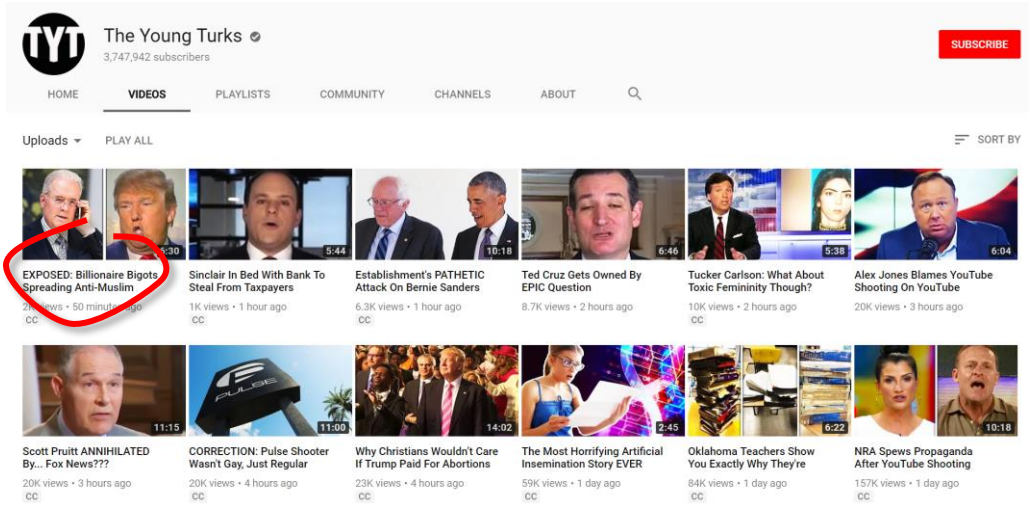


Figure 1.

In case you are logged into your Google account in your Chrome browser, please log off for this exercise (click on the top right of the YouTube page, etc). If you have any add-blocker installed on your Chrome browser, it's best to disable it for this.

Open the URL you want to scrape (in this case www.youtube.com/user/TheYoungTurks/videos) and open up the Web Scraper by right clicking on the page and select >**Inspect**<, or by keyboard shortcuts.

Note: A panel will open up. You can make it larger and smaller. Your **Inspect** panel might be set on the right-hand side by default. We will move it and dock it to the bottom of the window.

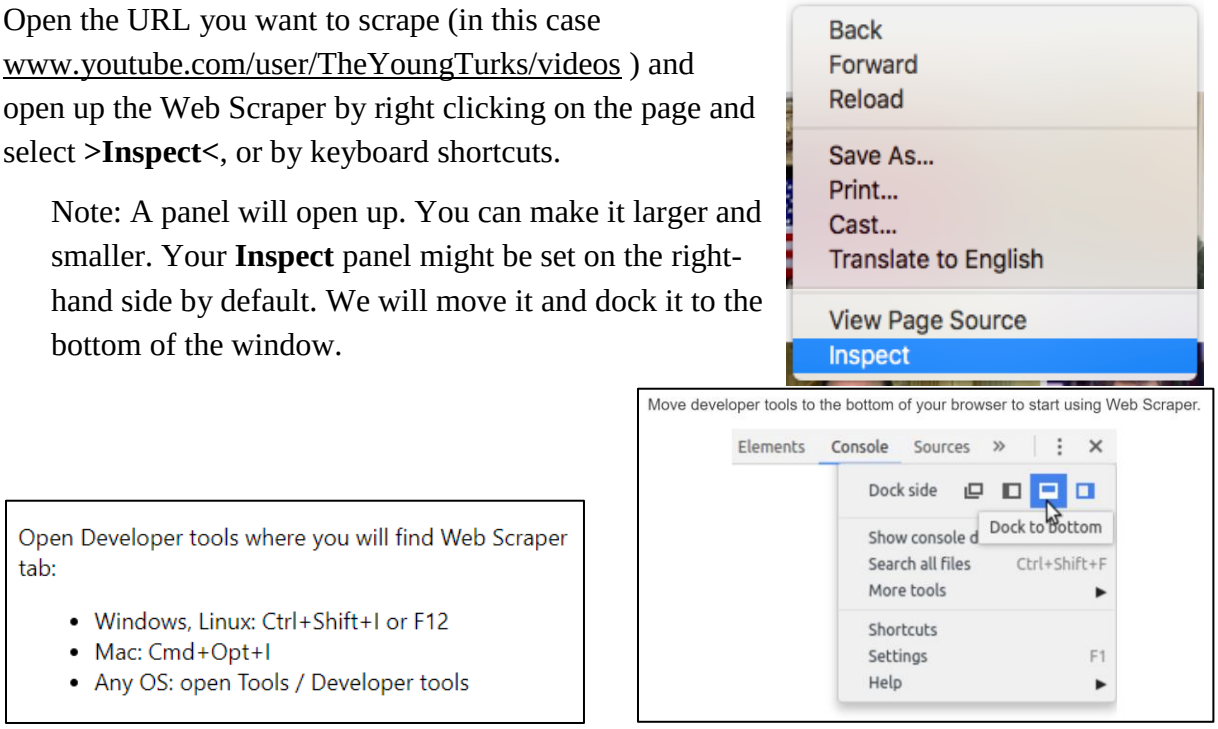


Figure 2

Then, click on the **Web Scraper** tab on the **Inspect** panel (a tab on top of the panel). Note: If you don't see it, you can click on the **" >> "** button on the panel to find the **Web Scraper** tab and follow the instructions there to move the panel to the bottom.

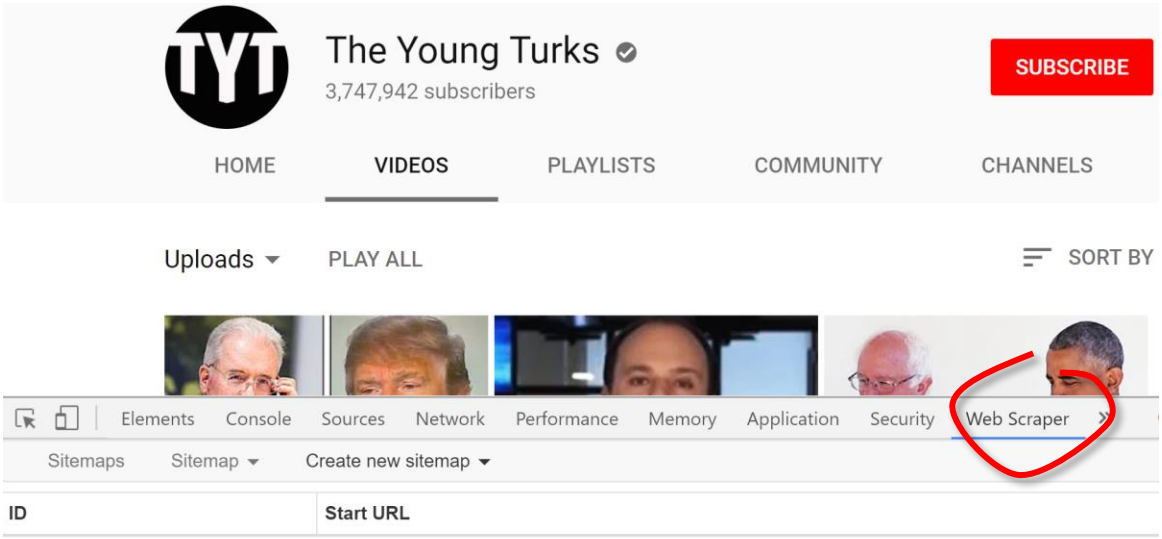


Figure 3.

Create a sitemap

Now we need to create a new sitemap, which literally is the “map” we develop of the “site” we want to scrape to guide the scraper. This is done by clicking the **>Create new sitemap<** tab (under the Web Scraper tab) and selecting **Create Sitemap**.

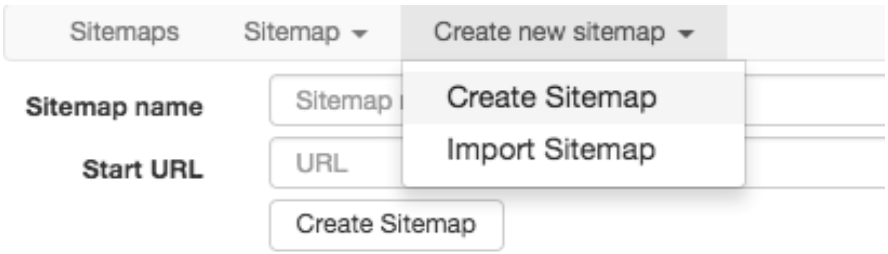


Figure 4.

Choose a **name** for our sitemap, we'll call it >youtube_example< here. Also, fill out the start page URL (copy-paste), in our case: https://www.youtube.com/user/TheYoungTurks/videos . Click >Create Sitemap<.

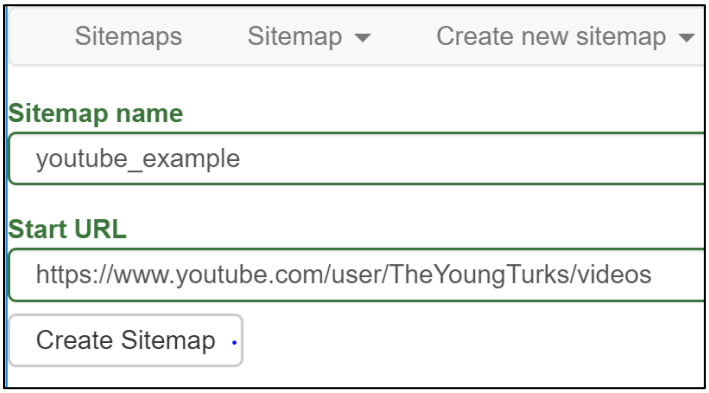


Figure 5.

Once we have created the sitemap, we'll get a new empty directory named `_root`.

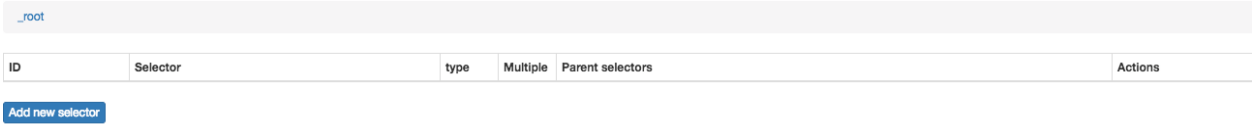


Figure 6.

Create list of videos

Now that we have an empty map, we need to tell the scraper what to select. We add a new **selector** for our scraper by clicking **Add new selector** button (see Figure 6).

We'll call our first selector, "featured" since they're the featured video of the channel. Switch the Type to be `>Link<` (we don't only want the Text, but the title text and the URL), then click on the >**Select<** button under Selector (see Figure 7 bottom).

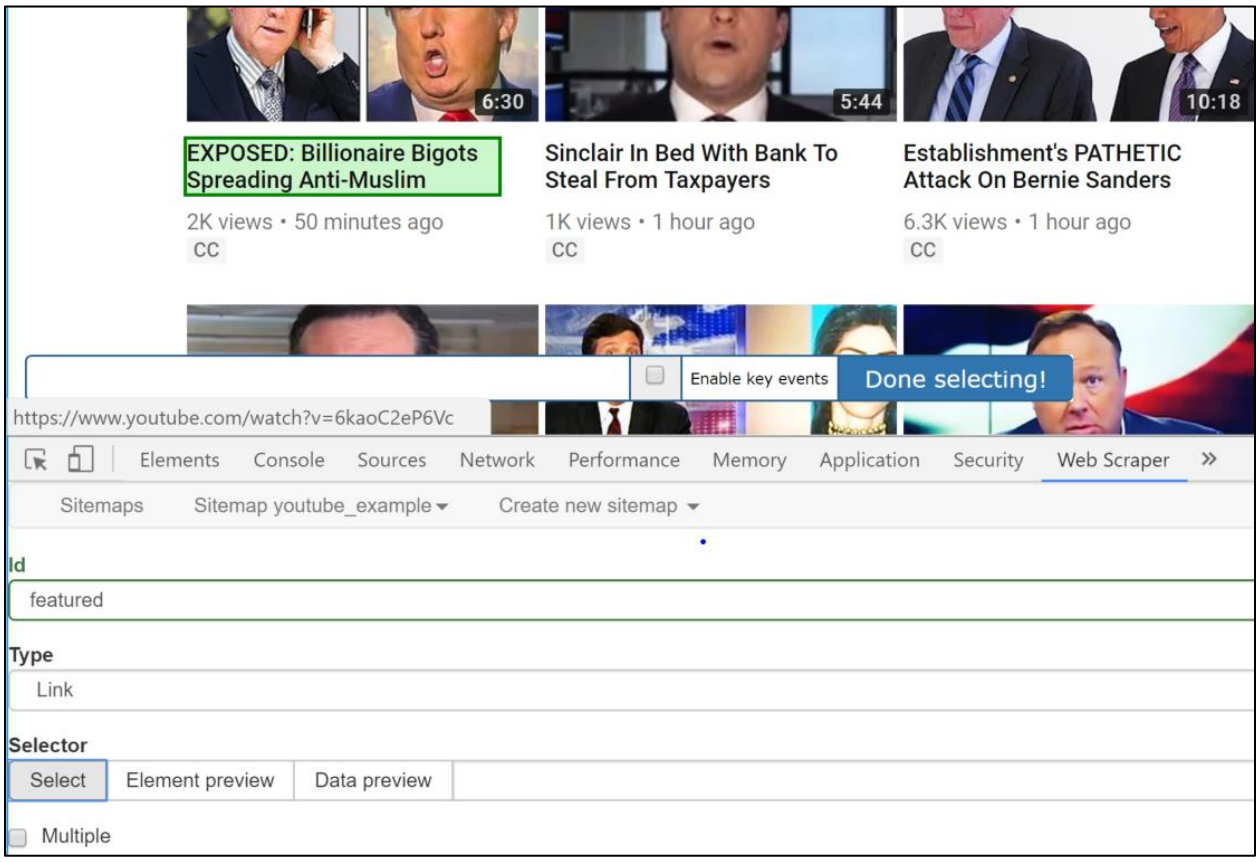


Figure 7.

Now we need to select what we want to scrape. Notice that when you hover your mouse over links from the webpage, they will change color (you might need to make the scraper control window a bit smaller). Move the mouse cursor over the first video until you selected the **video title ONLY** (for a more exensive collection see the accompanying pdf, that scrapes the title AND the number of views). Then click on the "green colored title" (see Figure 7). It should turn from green to red.

> Note: If you accidentally select something else, revert that choice by clicking again on the >**Select<** button under Selector.

Now, we have successfully located the first video for our scraper. We will now tell it to collect all similar links (all title links). This is where the magic happens, with the exception that it's not magic, but the web scraper taking advantage of the webpage architecture. Click on the second title, and the scraper will recognize all titles on the page as such, and color them red (Figure 8). The scraper realized that all of these elements have the same structure (which is embedded in the webpage) and concluded that we'd like to get all of them.

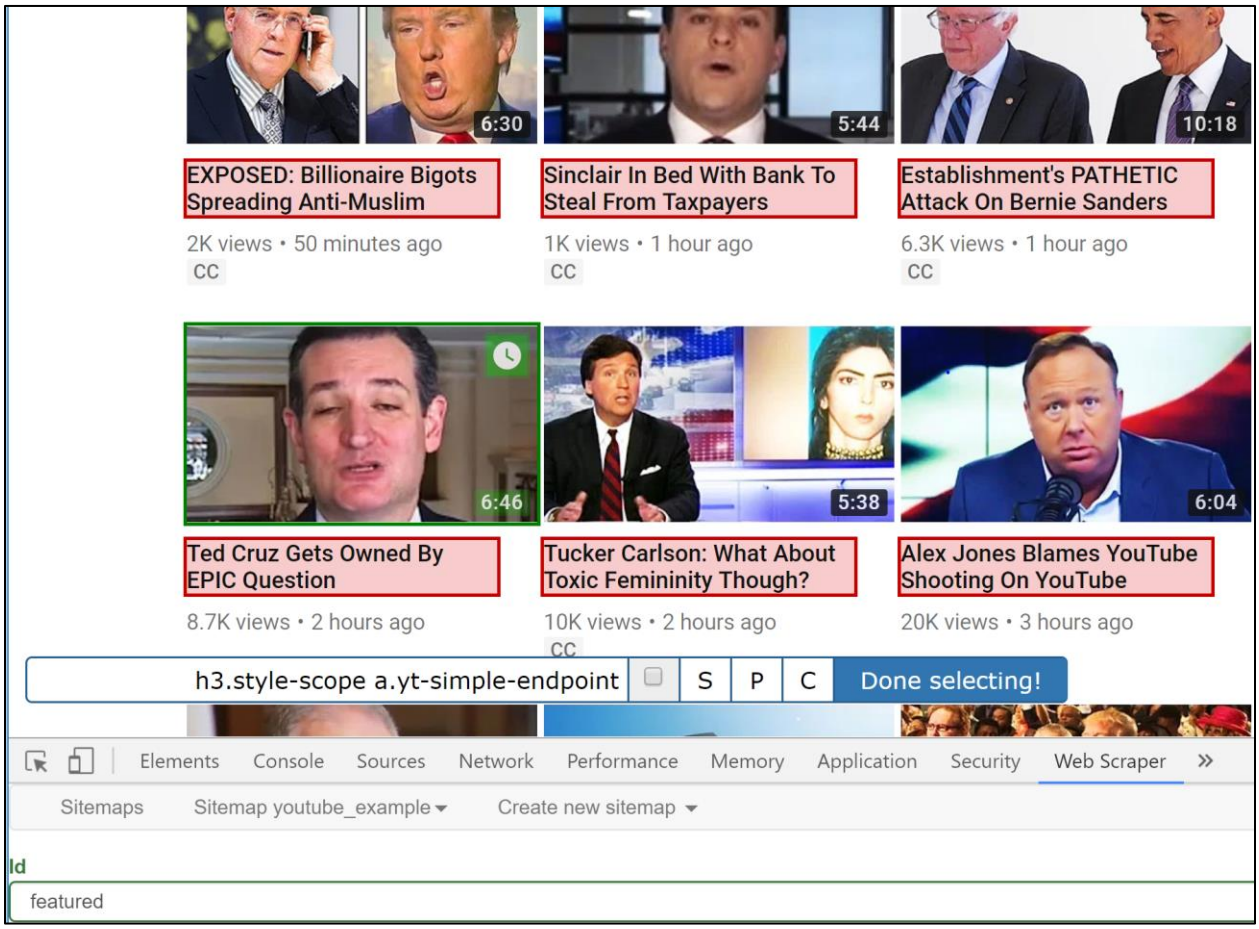


Figure 8.

Now you have to **save this selection** by clicking the **>Done selecting!<** button (see Figure 8).

Note: it only selected those titles that were visible at this point. If you would have scrolled down, more videos would have loaded. If you didn't, they had not loaded yet, and if it does not appear, you cannot scrape it (same as you could not "copy-paste" anything that is not there…). For our purposes, we are fine with the selection of the first videos.

Then click on the checkmark for >**Multiple**<, since we want the scraper to go through every video. Then click “Save selector” (Figure 9).

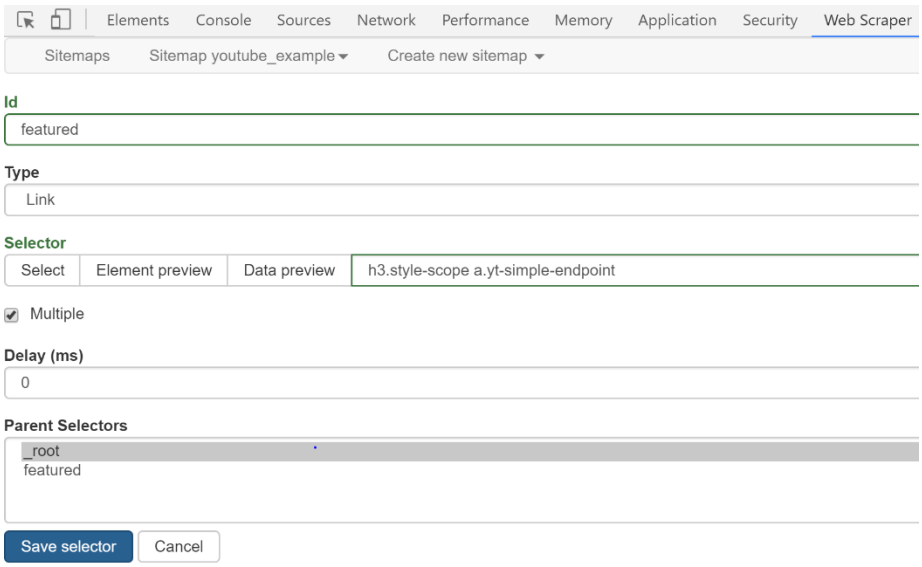


Figure 9.

We have now successfully created our first selector (see Figure 10)! It collects the “Links” (titlea and URLs) of the different videos featured in the channel.

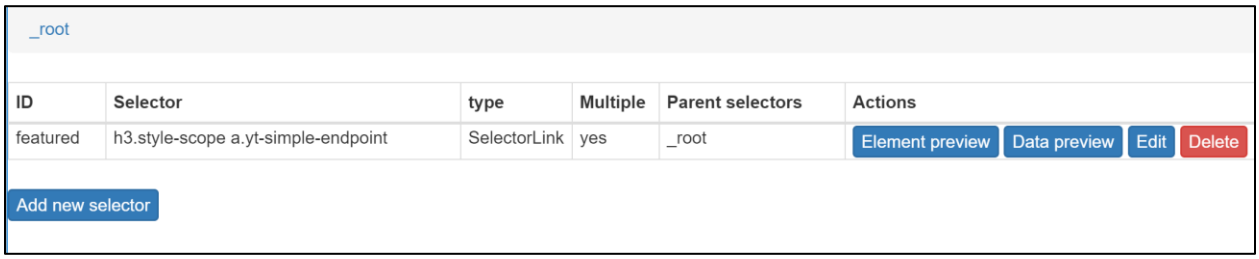


Figure 10

You can press “Data preview” (Figure 10) and you’ll preview your scraping harvest. If you don’t, go back to Figure 4, create a new sitemap, and start again.

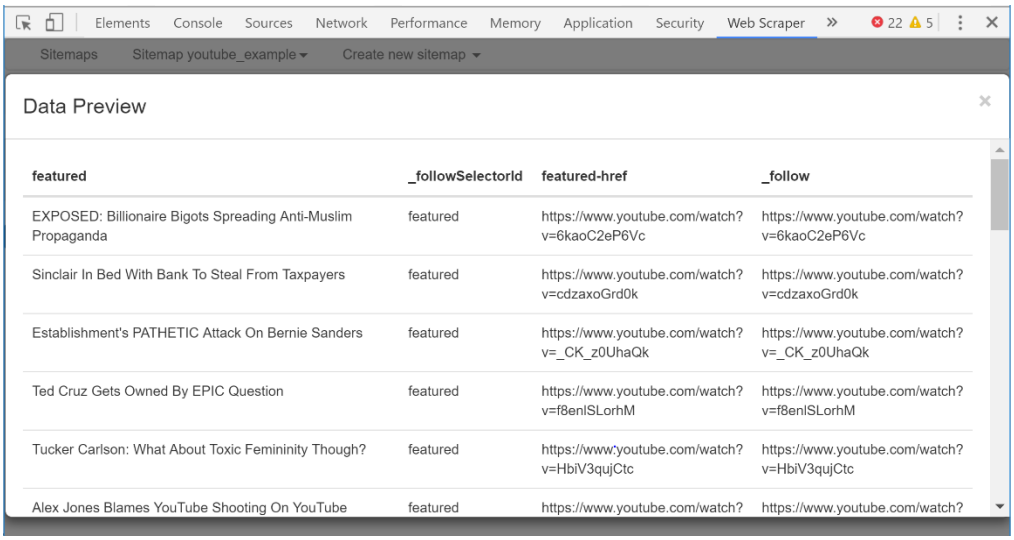


Figure 11

b. Recommended Videos

We also want the he **title** and **URL** of videos for all the "recommended videos" from **within each of the featured videos** (see Figure 12, highlighted area, which is inside one of the featured videos).

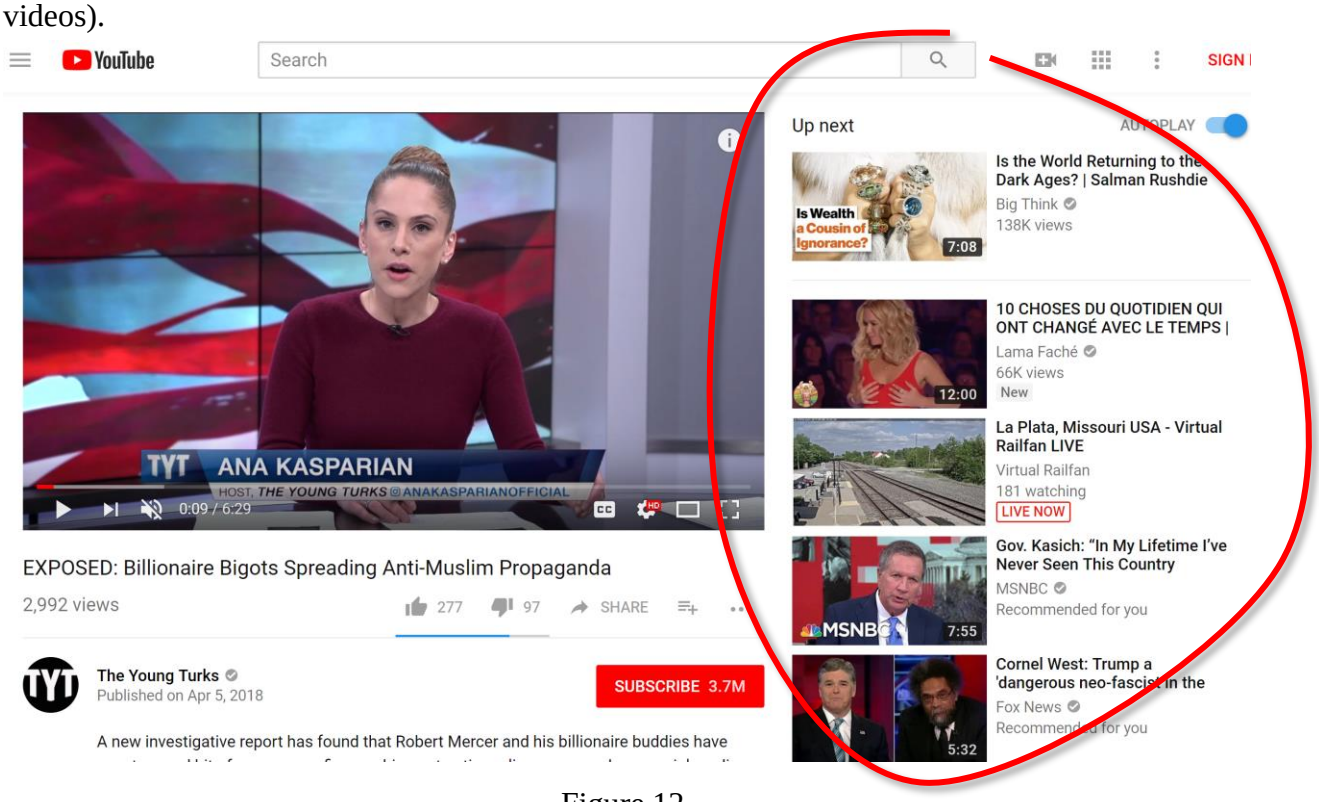


Figure 12.

Therefore, we have to go one level deeper, inside each of the collected URLs. Click on any video on the page to navigate to the page that plays a video (for example the first video from our list: for example, Figure 12 is the 1st video from Figure 7). Also move the selector one level deeper: click on the selector (row) called `featured` (see Figure 10), since it contains the link to the corresponding video player page. You are now within another sub-directory "**_root / featured**".

Note: stop the video if it start playing to prevent auto-play of a new video while coding.

We now create a selector similar to the `featured` selector we did earlier, but for the `recommended` videos.

- ✓ Add new selector
- ✓ Call Id: "recommended"
- ✓ Type: >Link<
- ✓ Selector: >Select<
- ✓ Hover over the first recommended video to mark it, then go to the second (get all the text this time, not only the title!)

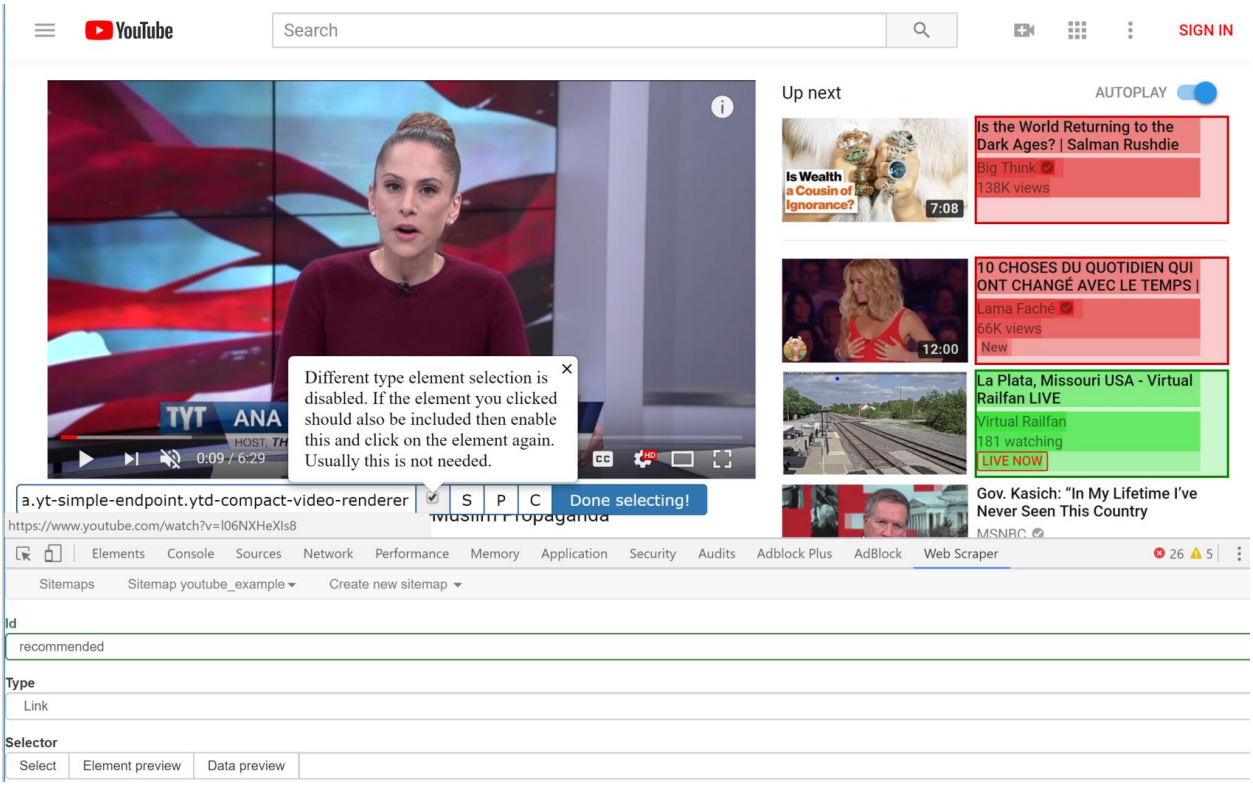


Figure 13.

Notice that the the "**Up next"** video is actually a different element on the page than the rest of the "recommended videos" (!). If you now go to the second recommended video, the scraper does not recognize both as the same, since the webpage treats both as different type elements. To overcome this, you need to click that little check box (see Figure 13). Now you can hover over and mark the second and once you hover and click on the third video, the scraper will recorgnize the remaining recommended videos as such, marking them red.

- ✓ Save with >**Done selecting!**<
- ✓ Remember the **Multiple** check box!
- ✓ Press >**Save selector**<

Note: If there is any advertisements on the site, please ignore them and do not select them!

You are done with the scraper for the recommended videos (Figure 14) and can now preview the data (will look similar to Figure 11: if not, go back up one level again, and repeat from Figure 12). In general: if something does not work, it is often easier (for such striaghtfoward codes) to simply redo everything. Might be quicker than debugging…

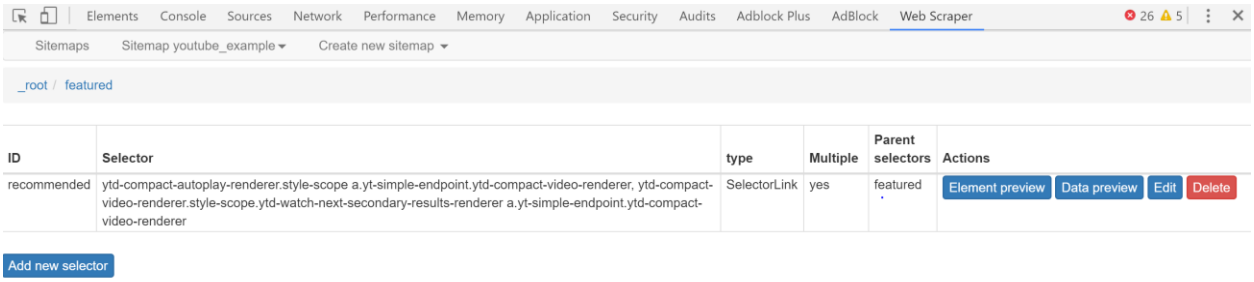


Figure 14.

We can check out a graphical representation of our scrapter by going to “Sitemap youtube_example” and clicking the >**selector graph**< button. One node will appear, called “_root”. Click on that (and subsequent) nodes, to expand the three branches that represent the flowchart of this algorithm.

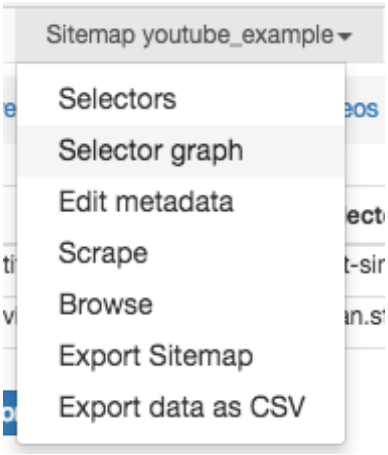


Figure 15.

3. Scrape

If everything looks good, we can move on to the scraping part. Start the scraper going under "Sitemap youtube_example" and click the >**Scrape**< button. Then click >**Start scraping**<.

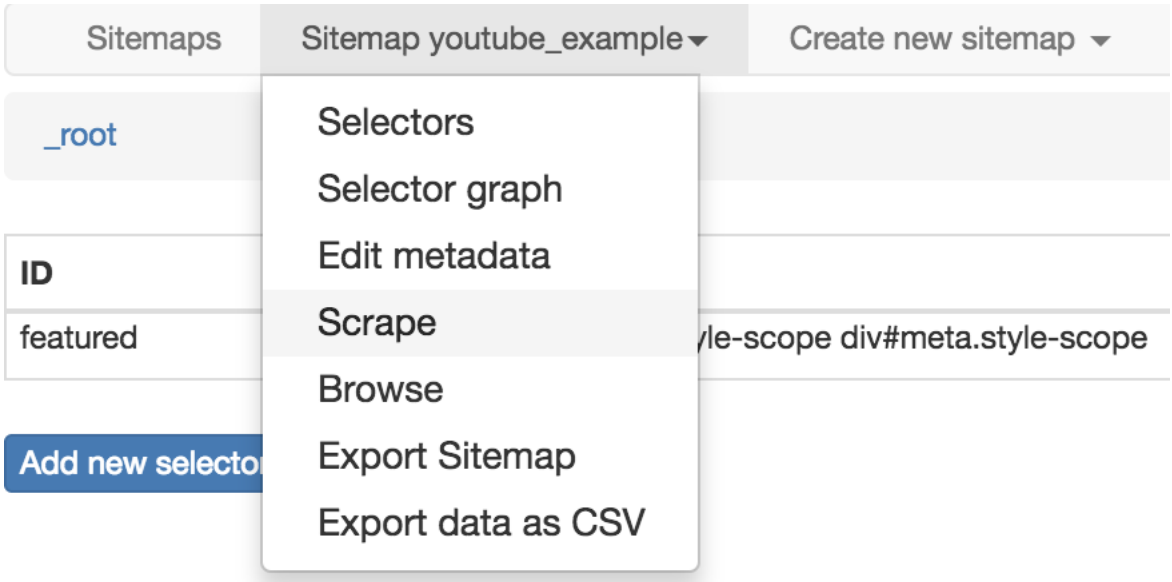
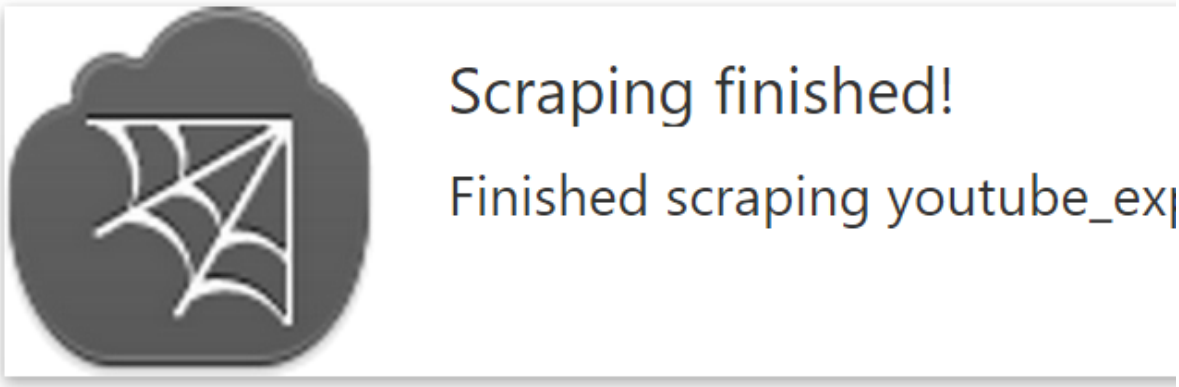


Figure 16.

A new window will pop-up and rattle through all the videos you are scraping. Good that you don't have to do this by hand! Here you see your algorithm doing the job for you. Scraping is done when the new window is closed. Press >**refresh**< (which will show up) to see the data.



4. Browse & Export & Check

After the scraping is done, we can output the data to a CSV file that we can open in MS **Excel**, Apple's iWork Numbers, Google Sheets, or LibreOffice spreadsheets, etc.

Click on **Export data as CSV** to enter the download page. Click on **Download now!** to download the file.

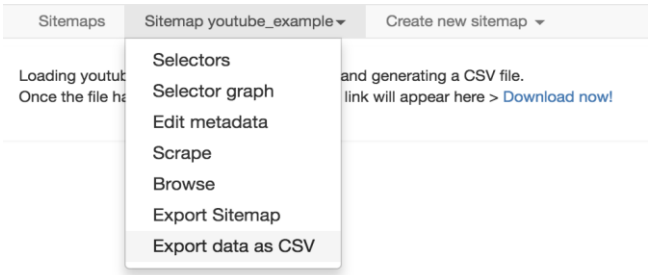


Figure 17.

Open your spreadsheet. Your spreadsheet should have six columns:

web-scraper-order | web-scraper-start-url | featured | featured-href | recommended | recommended-href

	A	B	C	D	E	F
1	web-scraper-order	web-scraper-start-url	featured	featured-href	recommended	recommended-href
2	1522986399-878	https://www.youtube.com/user/TheYoungTurks/videos	How Close Is America To Realizing MLK's Dream?	https://www.youtube.com/watch?v=0x76GbrZGjg	Trump's Unhinged Rally, Betsy	https://www.youtube.com/watch?v=yPx3q-E_6mM
3	1522986409-960	https://www.youtube.com/user/TheYoungTurks/videos	Oklahoma Teachers Show You Exactly Why They're Walking	https://www.youtube.com/watch?v=y98i2MgNzus	The Most Horrifying Artificial	https://www.youtube.com/watch?v=xjQ9D9uCPjw
4	1522986378-736	https://www.youtube.com/user/TheYoungTurks/videos	Shooting at YouTube HQ	https://www.youtube.com/watch?v=p1CjWuWQPNA	#1 Trump Leaker EXPOSED	https://www.youtube.com/watch?v=6WVNtQB5FP
5	1522986422-1070	https://www.youtube.com/user/TheYoungTurks/videos	Alex Jones Blames YouTube Shooting On YouTube	https://www.youtube.com/watch?v=wWn0YFBo-og	Scott Pruitt ANNIHILATED By...	https://www.youtube.com/watch?v=Py_SZt9el_E
6	1522986381-755	https://www.youtube.com/user/TheYoungTurks/videos	White Supremacy World With Tucker Carlson	https://www.youtube.com/watch?v=0m6Xz6s0EdI	Parkland Survivor Epically Takes	https://www.youtube.com/watch?v=Ggy3aYSamWI
7	1522986365-632	https://www.youtube.com/user/TheYoungTurks/videos	Howard Kurtz Scolds Reporter For Insufficient Politeness To T	https://www.youtube.com/watch?v=xqP6iwHlMu0	CORRECTION: Pulse Shooter	https://www.youtube.com/watch?v=vFCAFZGsGYs
8	1522986386-795	https://www.youtube.com/user/TheYoungTurks/videos	Trump Eyeing Up His Next Firing Victim	https://www.youtube.com/watch?v=BgACVXAljAc	Ken Bone: The Most Open-	https://www.youtube.com/watch?v=VRsc_4X2sJY
9	1522986430-1132	https://www.youtube.com/user/TheYoungTurks/videos	Establishment's PATHETIC Attack On Bernie Sanders	https://www.youtube.com/watch?v=_CK_z0UhaQk	Cop Who Shot 15-year-old	https://www.youtube.com/watch?v=mERUGHzws9
10	1522986404-924	https://www.youtube.com/user/TheYoungTurks/videos	New Details Emerge On Youtube Shooter's Motives	https://www.youtube.com/watch?v=X_9LCZPLhrU	CNBC-TV18 LIVE STREAM |	https://www.youtube.com/watch?v=RT1Yqxgj_MY
11	1522986368-657	https://www.youtube.com/user/TheYoungTurks/videos	Kushner Bragged About Sinclair Deal For Pro-Trump Coverag	https://www.youtube.com/watch?v=tzF1X57oTX8	Oklahoma Teachers Show You	https://www.youtube.com/watch?v=y98i2MgNzus
12	1522986389-804	https://www.youtube.com/user/TheYoungTurks/videos	Alison Hartson Ad EXPOSES Dianne Feinstein	https://www.youtube.com/watch?v=mfUQjmQ8Ck0	Republican Twitter Admits How	https://www.youtube.com/watch?v=0A7w7aj0jtk
13	1522986417-1031	https://www.youtube.com/user/TheYoungTurks/videos	CORRECTION: Pulse Shooter Wasn't Gay, Just Regular Terrori	https://www.youtube.com/watch?v=vFCAFZGsGYs	Michelle Williams On How Kerr	https://www.youtube.com/watch?v=b2rIQbCeawk
14	1522986396-862	https://www.youtube.com/user/TheYoungTurks/videos	What MLK Was Working On Before His Death	https://www.youtube.com/watch?v=tD6RzBeicKI	NRA Spews Propaganda After	https://www.youtube.com/watch?v=3k8LK0wpHzs
15	1522986362-614	https://www.youtube.com/user/TheYoungTurks/videos	Sinclair Issues Employee Gag Order After Trump Propaganda	https://www.youtube.com/watch?v=KNdFBDz0B3g	Ted Cruz Gets Owned By EPIC	https://www.youtube.com/watch?v=f8enlSLorhM
16	1522986407-946	https://www.youtube.com/user/TheYoungTurks/videos	NRA Spews Propaganda After YouTube Shooting	https://www.youtube.com/watch?v=3k8LK0wpHzs	Propaganda (DJ Pomeha) - You	https://www.youtube.com/watch?v=mCrrr-U4DpE
17	1522986386-781	https://www.youtube.com/user/TheYoungTurks/videos	Trump Eyeing Up His Next Firing Victim	https://www.youtube.com/watch?v=BgACVXAljAc	#1 Trump Leaker EXPOSED	https://www.youtube.com/watch?v=6WVNtQB5FP
18	1522986394-857	https://www.youtube.com/user/TheYoungTurks/videos	Alexandria Ocasio-Cortez Takes On Corporate Democrat Joe	https://www.youtube.com/watch?v=ORkJDZLIIKs	Tucker Carlson: What About	https://www.youtube.com/watch?v=HbiV3qujCtc
19	1522986383-777	https://www.youtube.com/user/TheYoungTurks/videos	DEA Arrests Pill-Pushing Pharmacists	https://www.youtube.com/watch?v=pMiqIgdcpb0	Judge Rules Mass Shooter	https://www.youtube.com/watch?v=GrOKVzcIqVM
20	1522986396-861	https://www.youtube.com/user/TheYoungTurks/videos	What MLK Was Working On Before His Death	https://www.youtube.com/watch?v=tD6RzBeicKI	How Martin Luther King Jr. Was	https://www.youtube.com/watch?v=EXwfUhyrBvA
21	1522986383-764	https://www.youtube.com/user/TheYoungTurks/videos	DEA Arrests Pill-Pushing Pharmacists	https://www.youtube.com/watch?v=pMiqIgdcpb0	Ted Cruz Gets Owned By EPIC	https://www.youtube.com/watch?v=f8enlSLorhM
22	1522986396-874	https://www.youtube.com/user/TheYoungTurks/videos	What MLK Was Working On Before His Death	https://www.youtube.com/watch?v=tD6RzBeicKI	Smithsonian Channel releases	https://www.youtube.com/watch?v=NbQV1xKFQm

Figure 18.

Congratulations! How many rows did you collect?

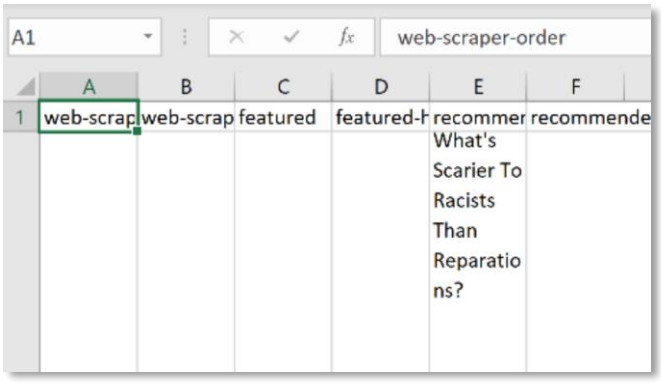
If you want, sort the rows alphabetically according to the column "feautred". Please search online how to do this, for example:

https://www.google.com/search?source=hp&q=Sorting+in+Excel

https://www.google.com/search?source=hp&q=Sorting+in+iWork+Numbers

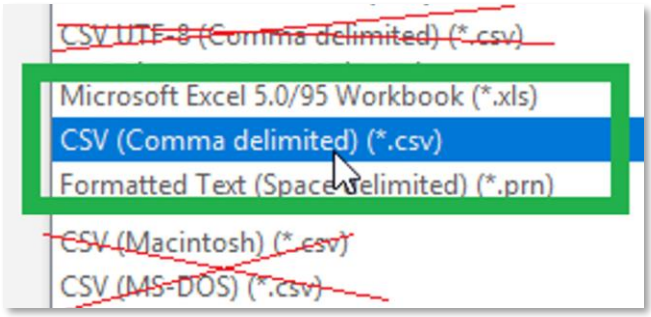
https://www.google.com/search?source=hp&q=sorting+in+Google+sheets

If your spreadsheet does not look as nicely organized as this one, the default in your spreadsheet software might be that the text in the cells is "wrapped".



Find out how to deal with this online, for example:
https://www.google.com/search?q=excel+wrap+text

Also, make sure you save the file as plain "CSV", when saving it to your computer ("save as"):



In general, as always when struggling with a computational task, simply do an online search where you describe your problem: almost certainly someone else already found a solution for it!