

Practical Web Scraping for Data-driven Social Science

Collecting data from online

Week 06

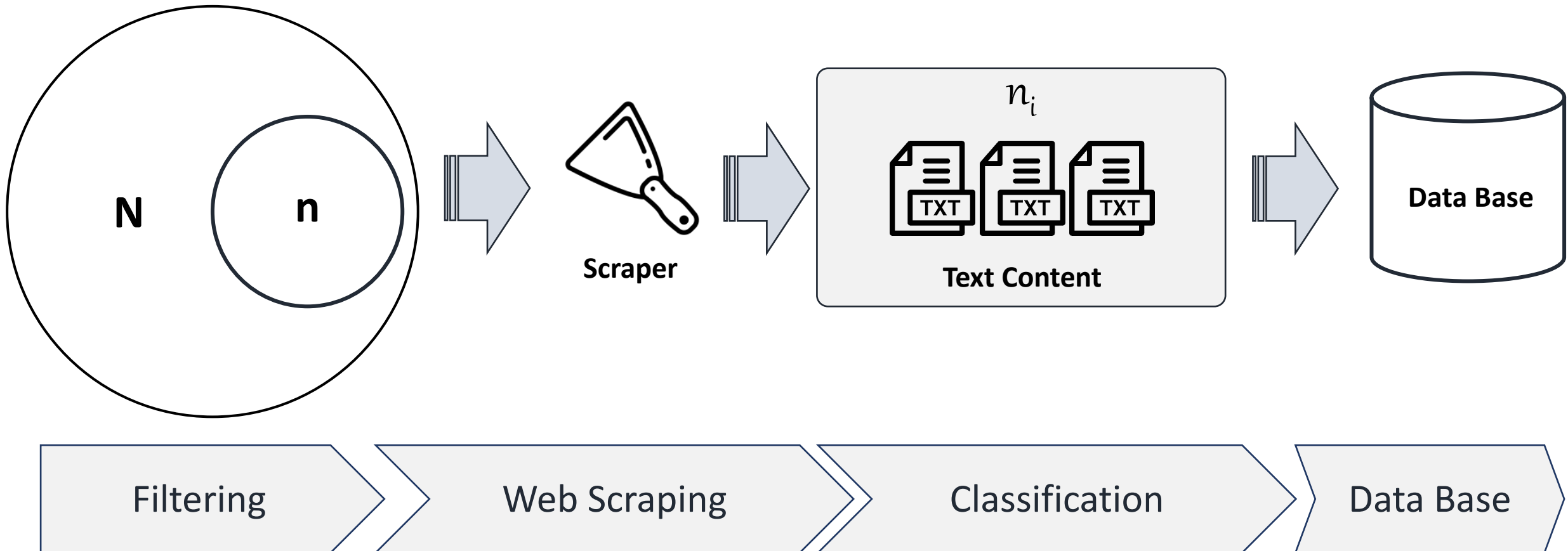
WEB SCRAPING



What is Web Scrapping and Why?

‘the construction of an agent to download, parse, and organize data from the web in an automated manner.’

<https://www.domain.com>



Lecture Task

If you think, "Somehow, I was able to get data from the website. I did it! "After this lecture, it will be a perfect score.

- Find a site that has the data you want.
- You know where your data is located on your website.
- Write code in R to get data.

Method 01: Semi Automatic, search the data you want

1. Search the website from google.com

site:damain.com search

The screenshot shows a Google search for "site:bbc.com putin ukraine invasion". The search bar and the search term are circled in red. Below the search bar, the "All" filter is selected. In the "Any time" dropdown menu, the "Custom range..." option is circled in red. To the right, a "Customised date range" dialog box is open, showing a calendar for January 2022. The "From" date is set to 1/1/2022 and the "To" date is set to 10/10/2022, both fields are circled in red. The "Go" button is also circled in red. The calendar shows the date 1/1/2022 is selected.

Google

site:bbc.com putin ukraine invasion

All News Videos Images Maps More Tools

Any time ▾ All results ▾

Any time
Past hour
Past 24 hours
Past week
Past month
Past year
Custom range...

Tracking the war with Russia - BBC N

forces have taken back more territory following Russ

Here are the latest developments:..

Invaded Ukraine and what does Putin

an leader's initial aim was to overrun Ukraine and di

government, ending for good its desire to join the Western defensive ...

Customised date range

From 1/1/2022

To 10/10/2022

Go

« January 2022 »

M	T	W	T	F	S	S
27	28	29	30	31	1	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16
17	18	19	20	21	22	23
24	25	26	27	28	29	30
31	1	2	3	4	5	6

The recommended search date range is

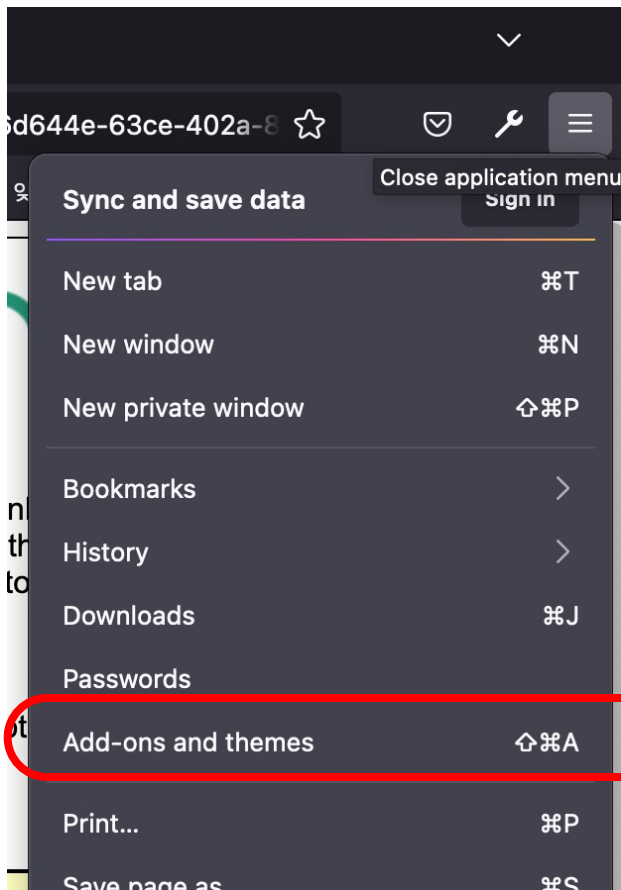
3month.

In general, google gives **lesser than 400 results.**

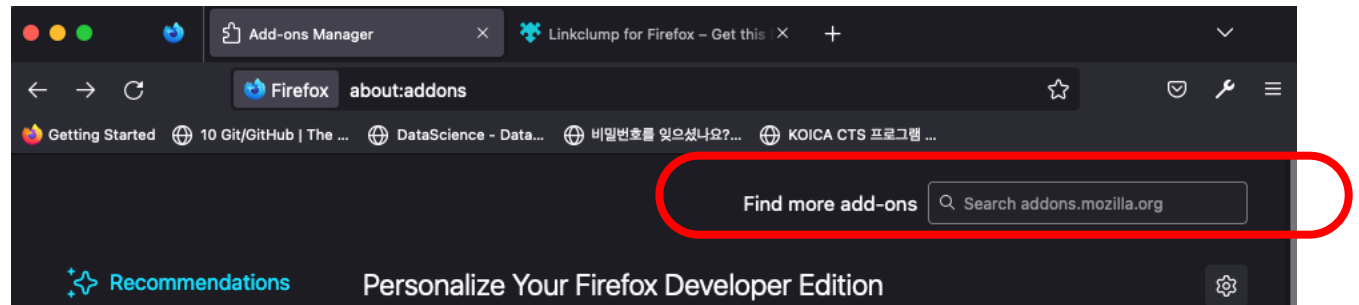
Method 01: Semi Automatic, search the data you want

1) Install Firefox Developer edition

<https://www.mozilla.org/en-US/firefox/developer/>



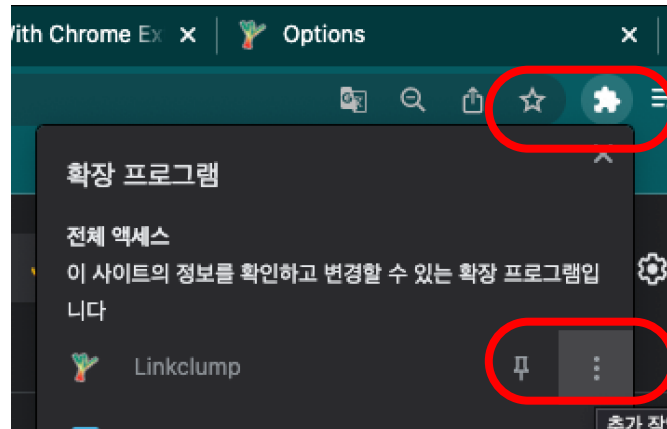
2) Click Add-ons and themes



2) Search “Linkclump” and **install**.

Method 01: Semi Automatic, search the data you want

- You can do it with Google Chrome, but I recommend FDE.
 - 1) You need to download a semi-automatic URL selector.
 - 2) Visit the following “**Chrome web store**” link.
<https://chrome.google.com/webstore/category/extensions>
 - 3) Search “Linkclump” and **install**.



4) Chrome Extension logo

5) Click the “additional work” and select “option”



Method 01: Semi Automatic, search the data you want

- Scrape the URL manually

The screenshot shows a configuration window for a web scraping tool. It is divided into three main sections: Activation, Action, and Advanced Options. The 'Activation' section is circled in pink and contains settings for the mouse button (Left), a key (z), and the selection box color (orange). The 'Action' section has four radio button options: 'Opened in a New Window', 'Opened as New Tabs', 'Bookmarked', and 'Copied to clipboard'. The 'Copied to clipboard' option is selected and circled in pink. The 'Advanced Options' section contains several settings: 'smart select' is set to 'on', 'filter links' is set to 'exclude links with words', and the 'copy format' dropdown menu is open, showing options like 'URLs with titles', 'URLS only' (which is selected and highlighted in blue), 'titles only', 'as link HTML', 'as list link HTML', and 'as Markdown'. A pink line also circles the 'filter links' and 'copy format' settings. At the bottom left is a 'Cancel' button. A small tooltip on the right says 'format of the links saved to the clipboard'.

Activation
to activate the selection box you should press and hold...

Mouse button: Left
+ key: z
Selection box color: orange

Action
Selected links should be...

☐ Opened in a New Window
☐ Opened as New Tabs
☐ Bookmarked
☒ Copied to clipboard

Advanced Options
Available options for the action...

smart select: on
filter links: exclude links with words
copy format: URLS only (selected)
titles only
as link HTML
as list link HTML
as Markdown

block repeat links in selection
reverse order

Cancel

format of the links saved to the clipboard

7) Leave as default

8) Select "Copied to clipboard"

9) Select filter links as "exclude links with words"

10) Select "URLS only" from the copy format

Method 01: Semi Automatic, search the data you want

Google site:bbc.com putin ukraine invasion

約 12,100 件 (0.51 秒)

<https://www.bbc.com/news/world-europe-60506682>
Ukraine in maps: Tracking the war with Russia - BBC News
6 日前 — Russia invaded Ukraine on 24 February, but Ukrainian forces retook large areas around Kyiv in early April after Russia abandoned its push ...

<https://www.bbc.com/news/world-europe-56720589>
Why has Russia invaded Ukraine and what does Putin want?
2022/05/09 — The Russian leader's initial aim was to overrun Ukraine and depose its government, ending for good its desire to join the Western defensive ...

<https://www.bbc.com/news/world-europe-63160354>
Putin's dream of Russian victory slips away in Ukraine - BBC
4 日前 — Like many, the political analyst believes that Mr Putin was caught completely off-guard by strong Western support for Kyiv, as well as Ukraine's ...

<https://www.bbc.com/news/world-60525350>
War in Ukraine - BBC News
Missiles strike Ukraine cities with central Kyiv hit. Vladimir Putin confirms Russia is behind the attacks, which come after an explosion targeted a key bridge ...

<https://www.bbc.com/news/world-europe-61990495>
Russia invasion: Putin still wants to take most of Ukraine - US
2022/06/30 — Russian President Vladimir Putin still wants to capture most of Ukraine, US intelligence agencies believe.

11) Press “Z” from the keyboard and drag the search results with the “left” click the mouse.

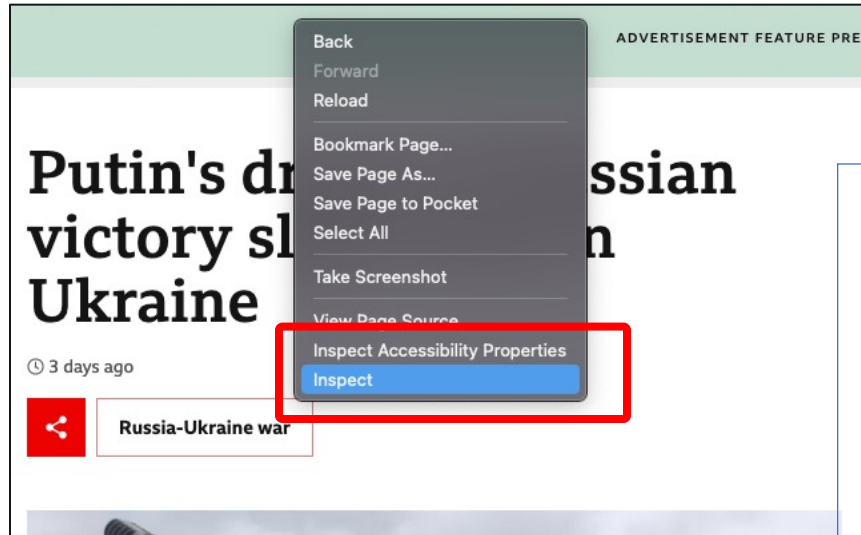
** This will save automatically to the clipboard*

	A	B	C	D	E	F	G	H
1	url							
2	https://www.bbc.com/news/world-europe-60506682							
3	https://www.bbc.com/news/world-europe-56720589							
4	https://www.bbc.com/news/world-europe-63160354							
5	https://www.bbc.com/news/world-60525350							
6	https://www.bbc.com/news/world-europe-61990495							
7	https://www.bbc.com/news/world-europe-62879367							
8	https://www.bbc.com/news/world-europe-63000034							
9	https://www.bbc.com/news/world-europe-62520743							
10	https://www.bbc.com/news/world-europe-62907923							
11	https://www.bbc.com/news/world-europe-61077648							
12	https://www.bbc.com/news/world-europe-60125659							
13	https://www.bbc.com/news/world-europe-61377886							
14	https://www.bbc.com/news/world-europe-62331061							
15	https://www.bbc.com/news/world-europe-61674469							
16	https://www.bbc.com/news/world-europe-60938544							
17	https://www.bbc.com/news/world-europe-61359228							
18	https://www.bbc.com/news/world-europe-63552630							

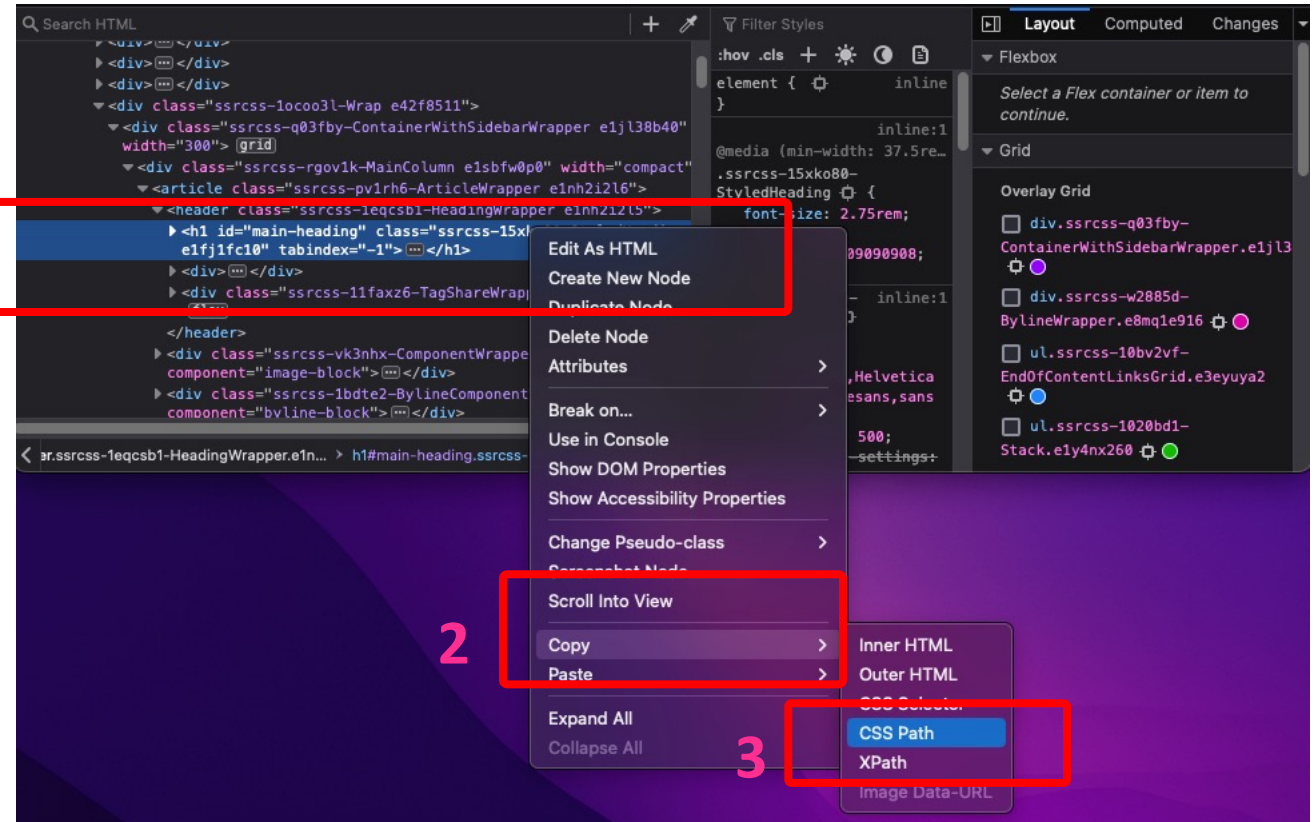
12) Open excel, create the “url” cell, and paste it above the "url" cell.

CSS, Xpath and Cording Practice: Web Scraping

Search the location of actual data



1) Role over to the target to inspect, click the “right” mouse button, and select “inspect”.



2) Move your mouse to the highlighted area and click the “right” mouse and select “copy”, and select either option of “CSS path” or “Xpath.”

Next Week

Please bring your laptop.

Week 7: Text mining fundamentals

Text manipulation in R, Quanteda grammar basics, and frequency tests

- What is data mining in a practical sense?
- Narrative Analysis
- Metaphor Analysis