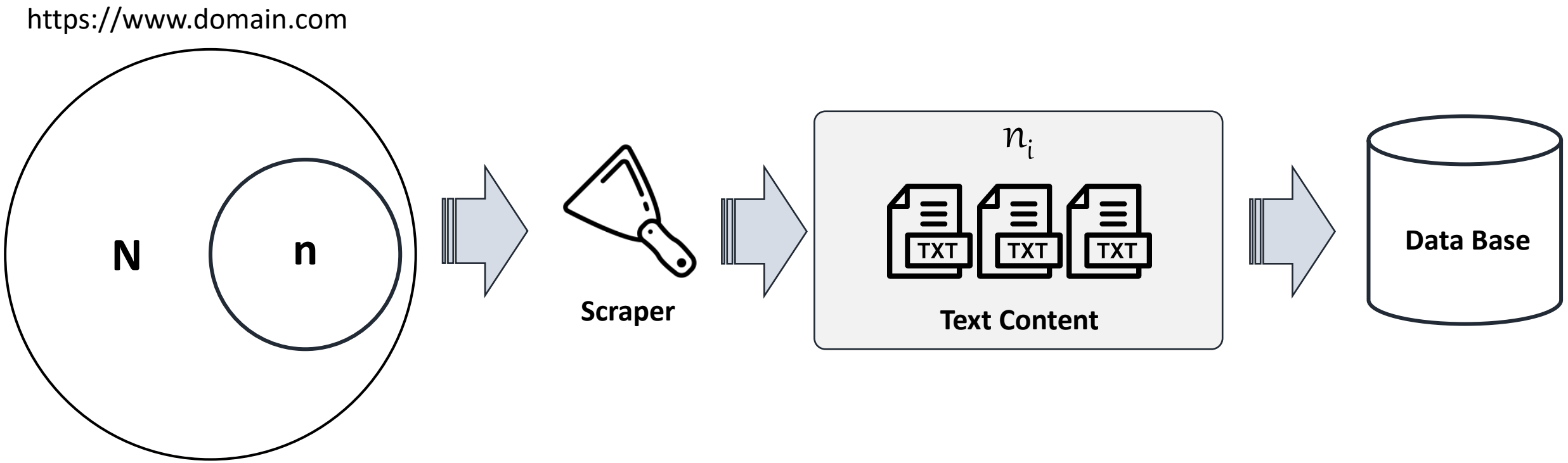
Practical Web Scraping for Data-driven Social Science

Collecting data from online

Week 06

What is Web Scraping and Why?

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



Filtering → Web Scraping → Classification → Data Base

Lecture Task

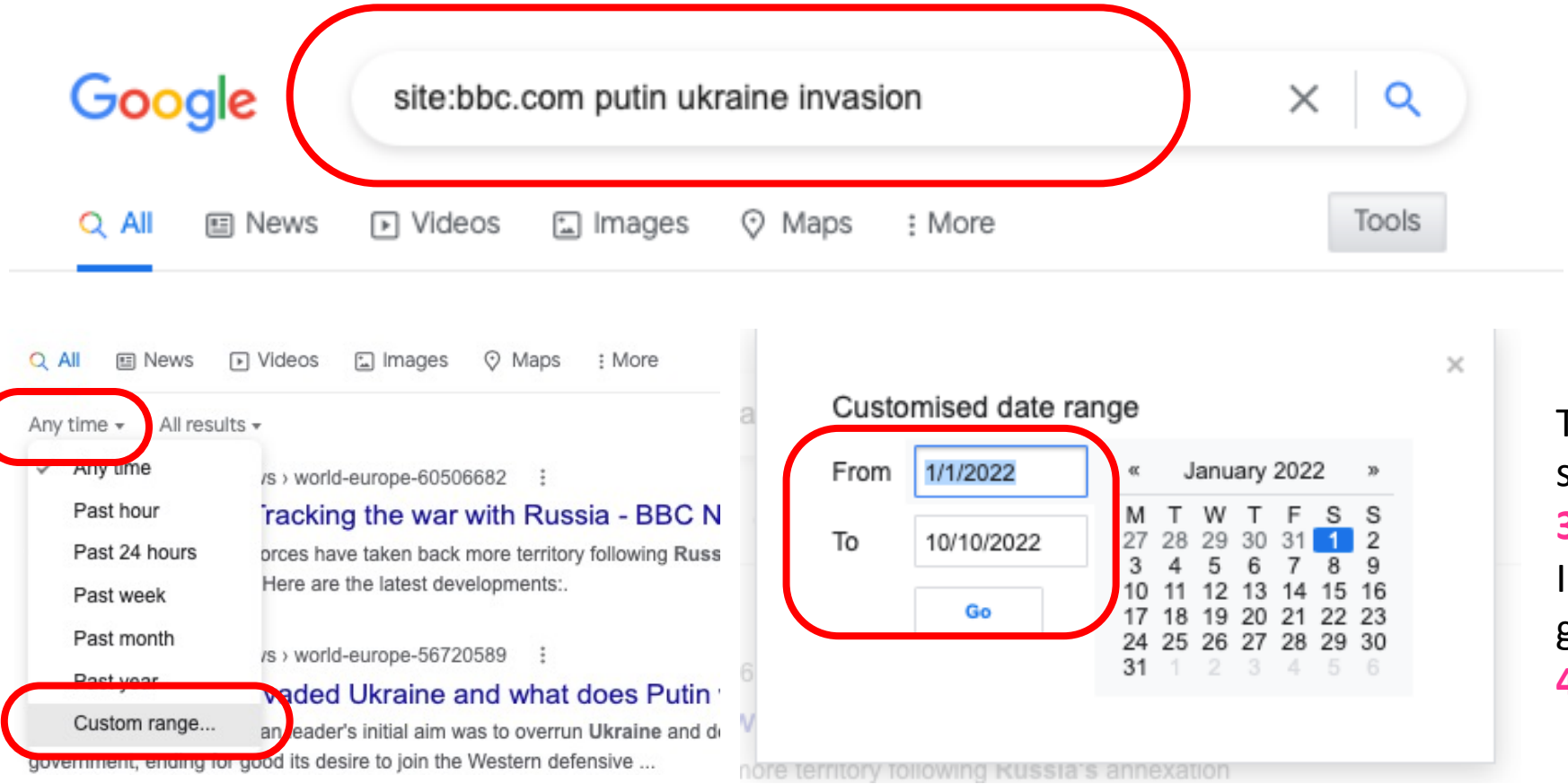
If you think, "Somehow, I was able to get data from the website. I did it! "After this lecture, it will be a perfect score.

- 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

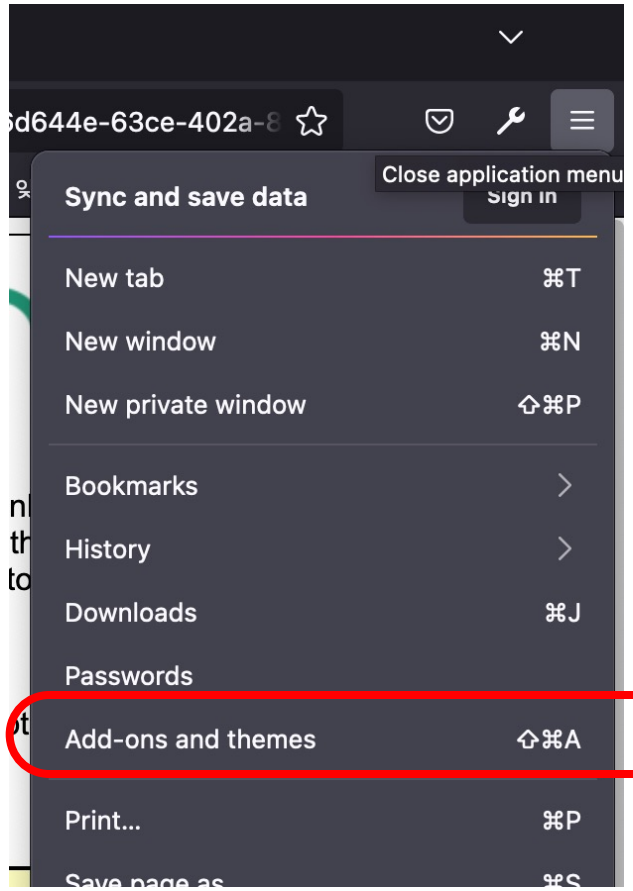


The recommended search date range is **3month**.
In general, google gives **lesser than 400 results.**

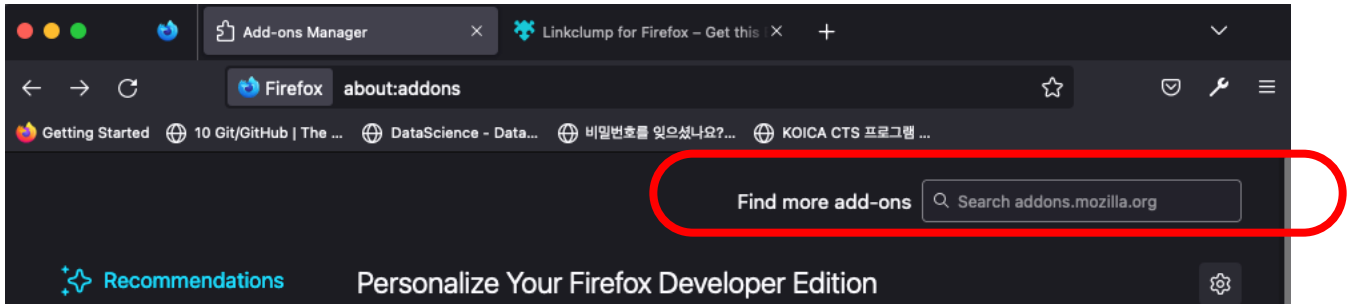
Method 01: Semi Automatic, search the data you want

1) Install Firefox Developer edition

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



2) Click Add-ons and themes

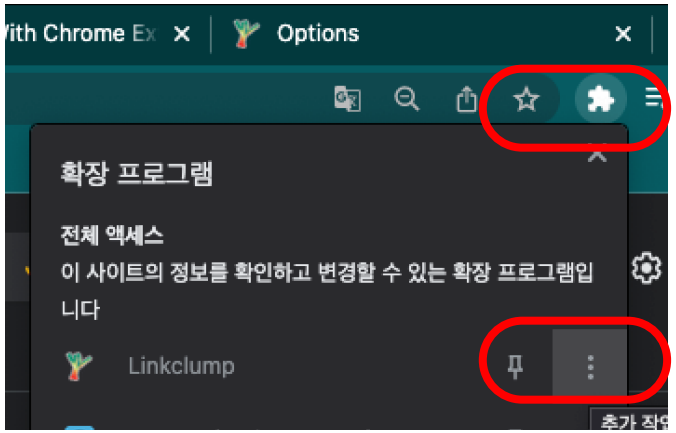


2) Search “Linkclump” and **install.**

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.**



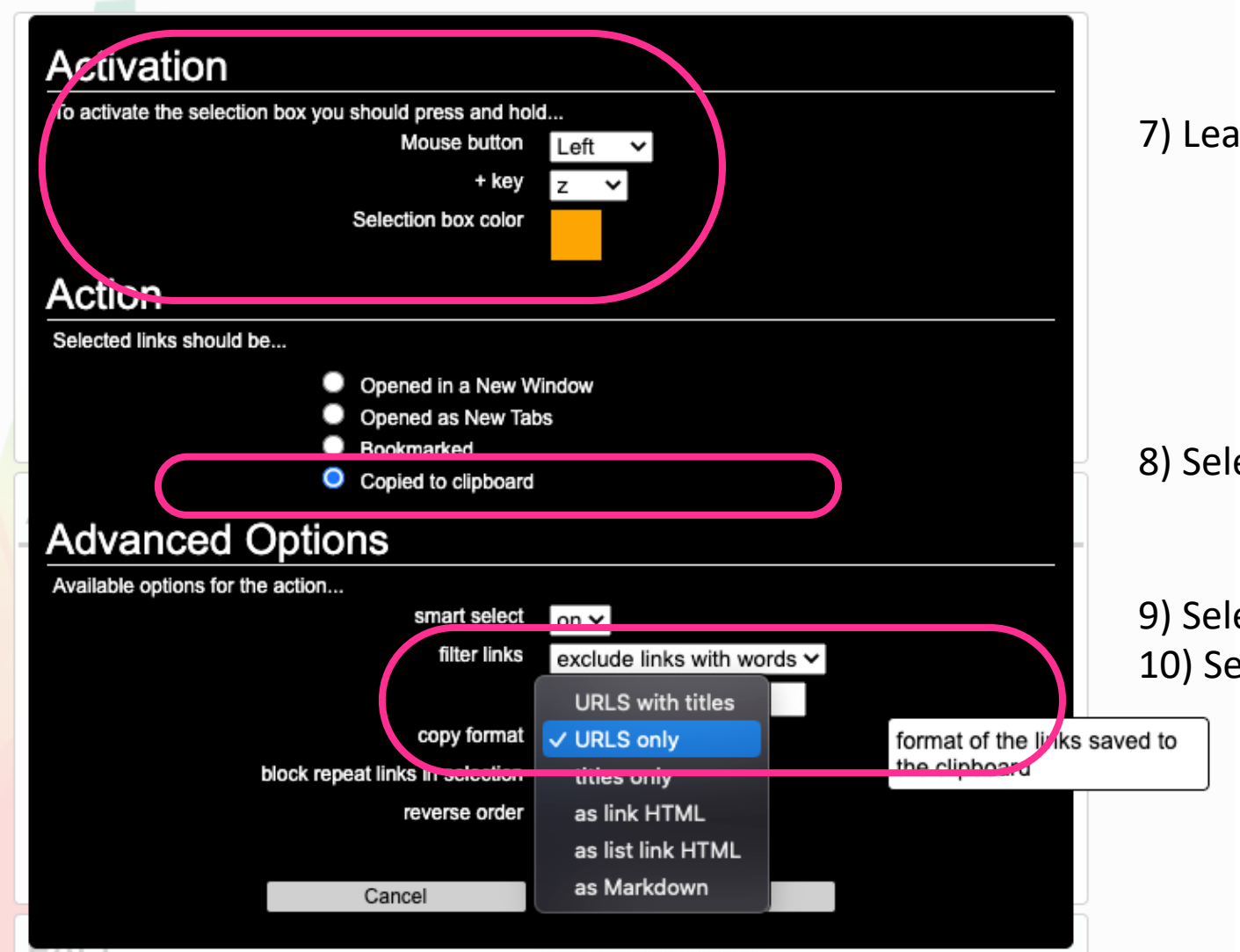
4) Chrome Extension logo

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

6) Click “Edit”

Method 01: Semi Automatic, search the data you want

- Scrape the URL manually

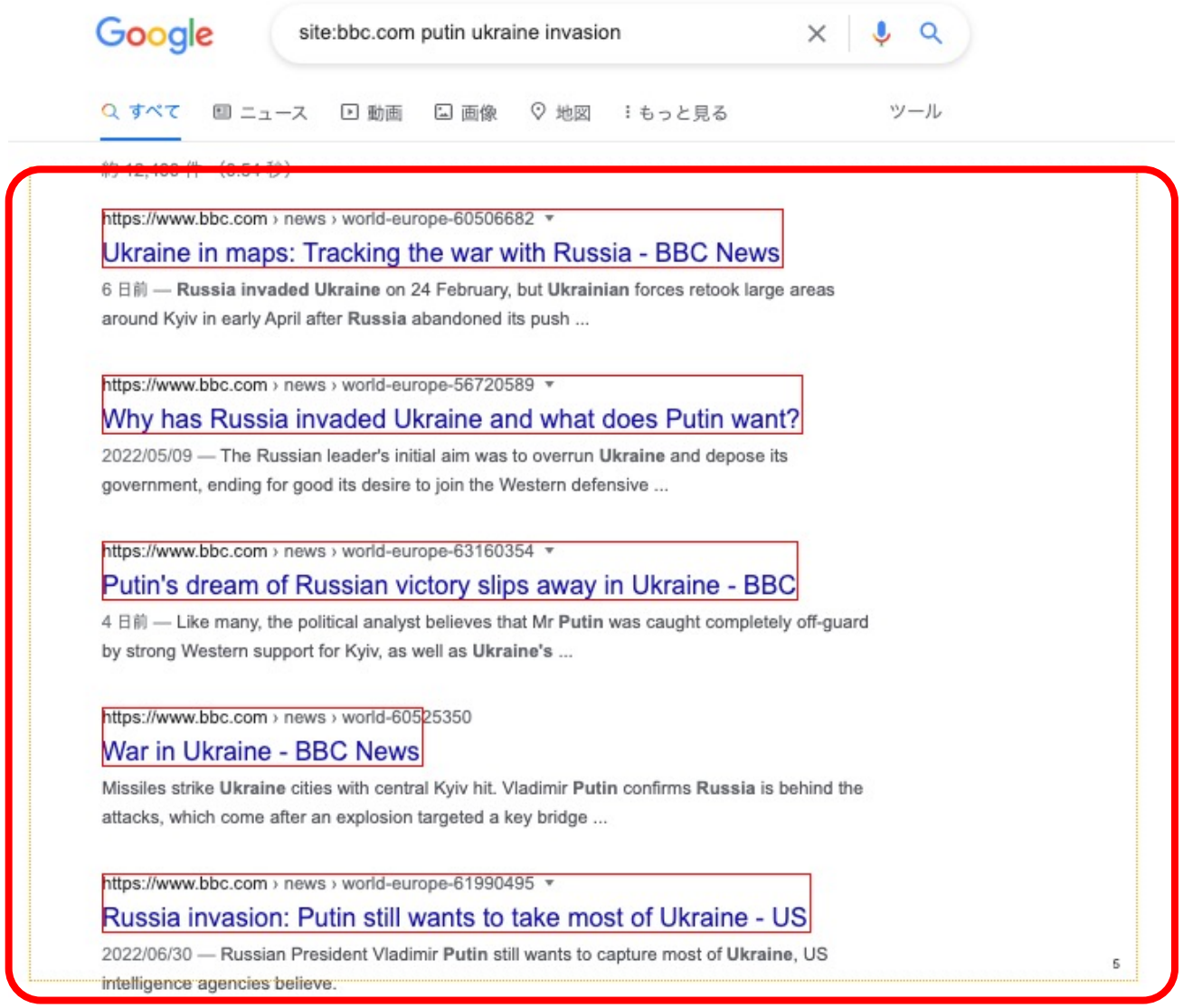


7) Leave as default

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



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

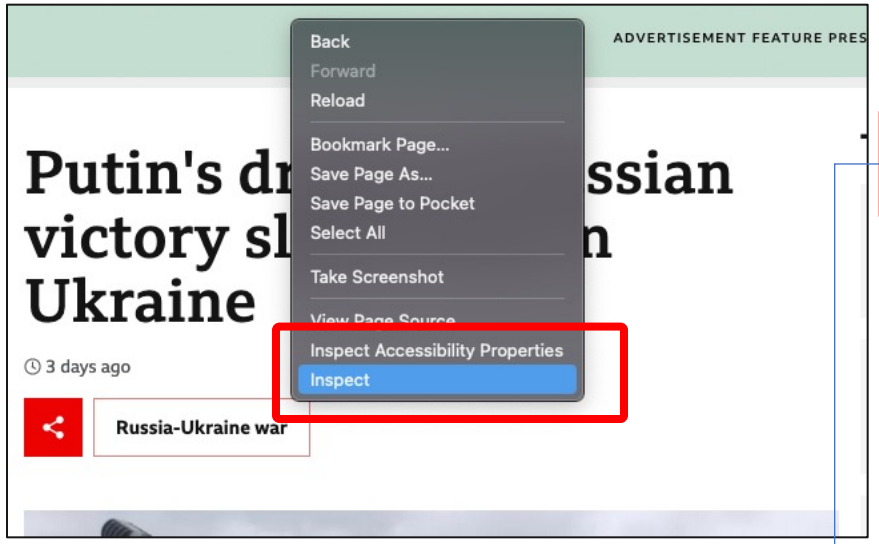
** This will save automatically to the clipboard*

	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							

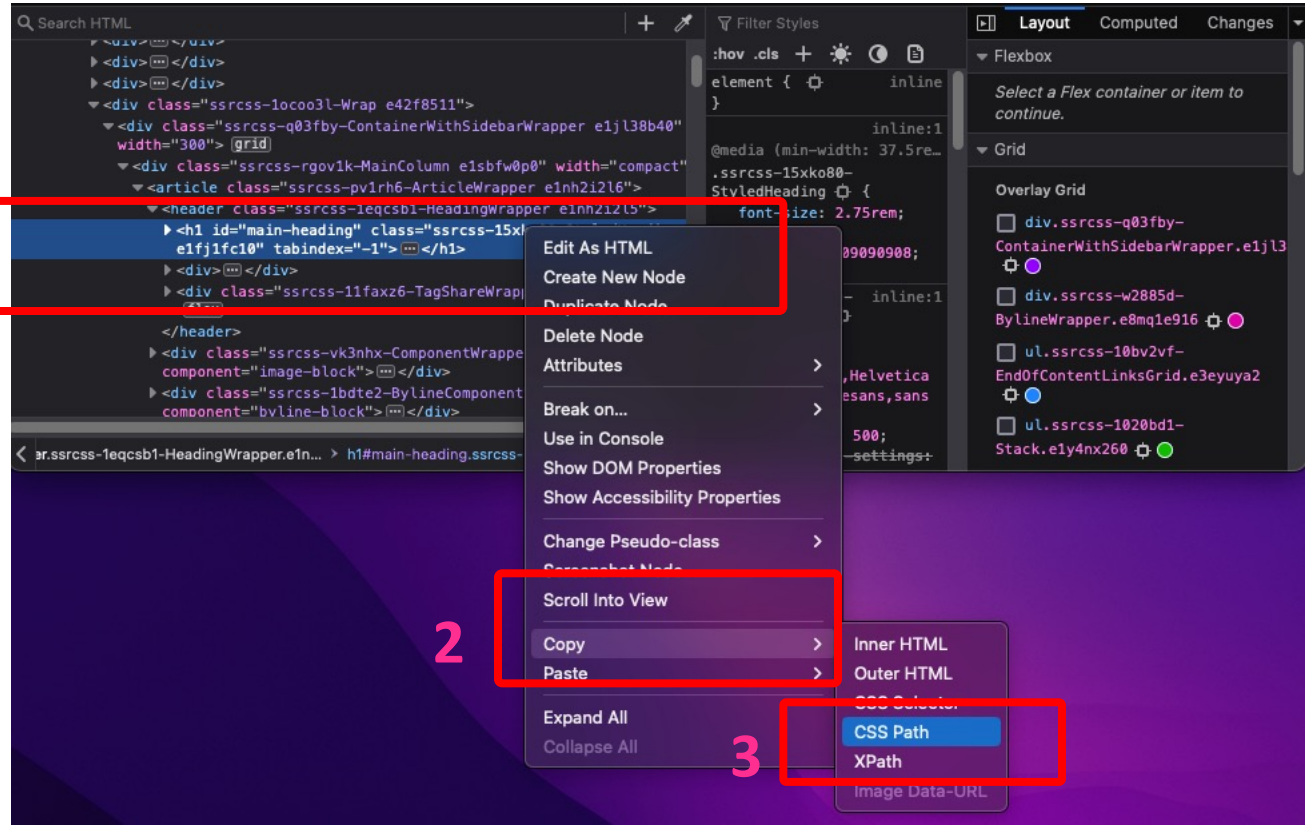
12) Open excel, create the "url" cell, and paste it above the "url" cell.

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