



Application of Big Data in Social Science

week 3

Heidi Hyeseung Choi
Fall 2022
HYSIS
heidichoi@hanyang.ac.kr

Announcements

9.1	1	Introduction
9.8	2	Web Scraping
9.15	3	
9.22	4	Natural Language Processing
9.29	5	
10.6	6	Text Analysis (recorded lecture on week 7)
10.13	7	
10.20	8	Mid term exam (as school schedule)
10.27	9	Social Network Analysis
11.3	10	
11.10	11	Machine Learning: Supervised Learning
11.17	12	
11.24	13	Machine Learning: Unsupervised Learning
12.1	14	
12.8	15	Data Visualization
12.15	16	Final Exam

In today's class!

1. Scrap multiple pages
2. Try out selenium
3. Scrap google news!



Recap from last week: web scraping

1. Retrieve HTML data from a domain name
 - `requests.get("webpage")`
2. Parse that data for target information using BeautifulSoup
 - make a soup first, with html parser
3. Store the target information
 - `find_all()` or `find()`
 - `get_text()`
4. make it into dataframe using Pandas
 - `pd.DataFrame(data= variable_name, columns = ['name of the column']`
 - combine dataframes horizontally using `pd.concat([data1, data2], axis = 1)`
5. export df to excel
 - `df_name.to_excel("document_name.xlsx")`





Some Python refreshers

- `import`
- `for` loops
- f-strings
- Concatenation of pandas library

importing libraries

- In Python, if you want to use a certain module, you need to ‘import’ that specific library. eg) pandas, BeautifulSoup, numpy, requests etc.
- It is the first thing that we do, when we start coding. This means you have to ‘import’ necessary libraries every time we launch jupyter notebook.

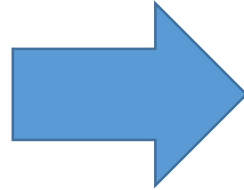
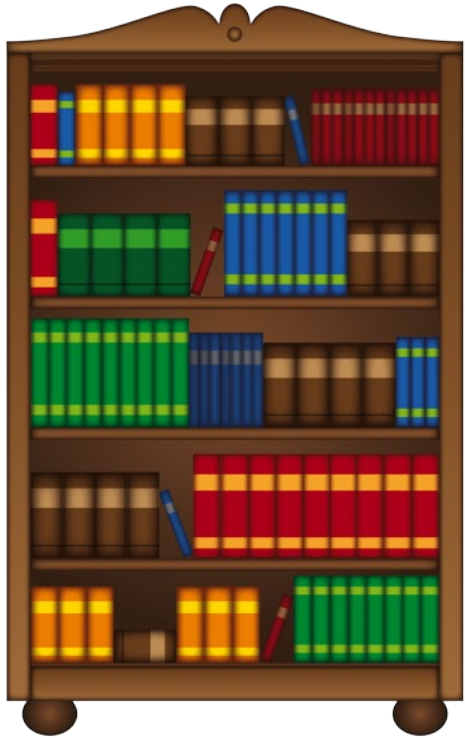
```
>>> from .... import ... as ...
```

```
>>> import ...
```

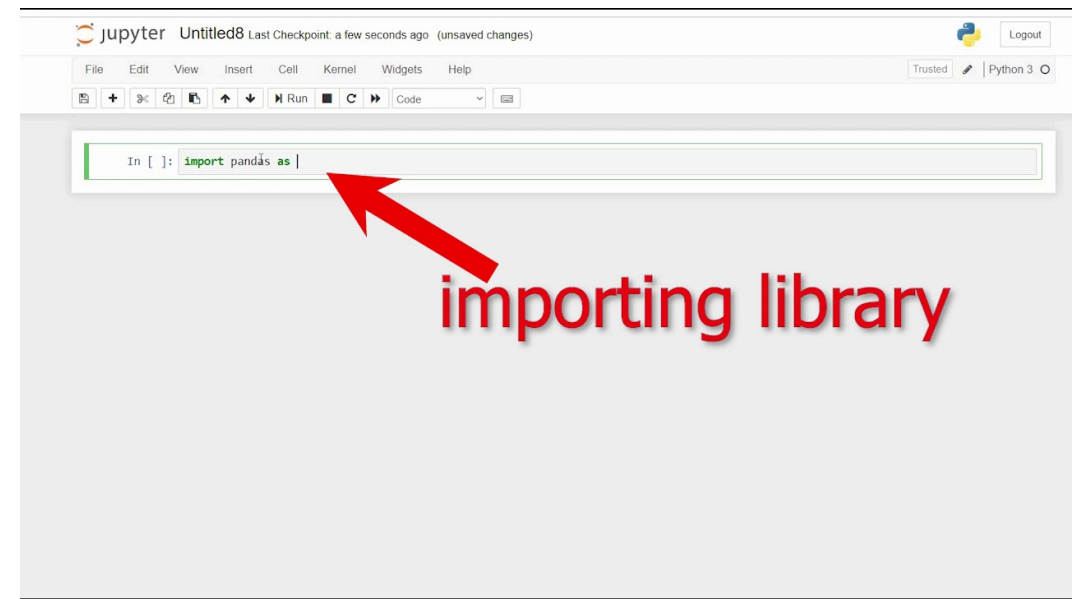


Python library

Python Standard Library



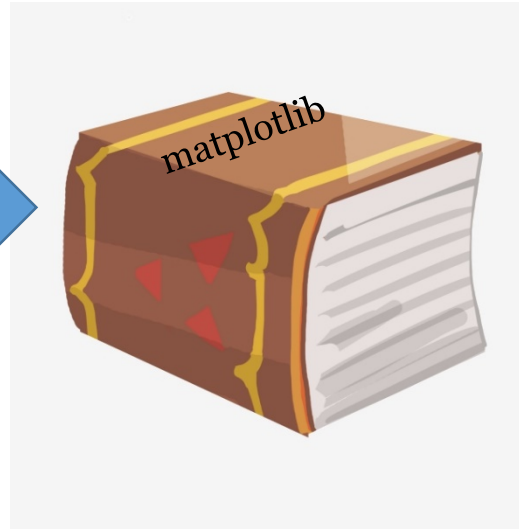
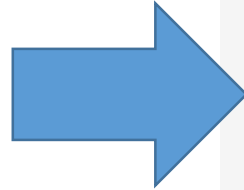
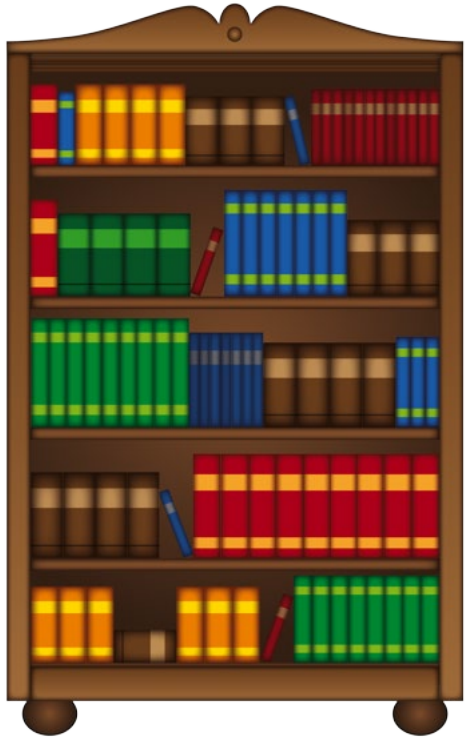
```
>>> import requests  
>>> import pandas  
>>> import BeautifulSoup
```



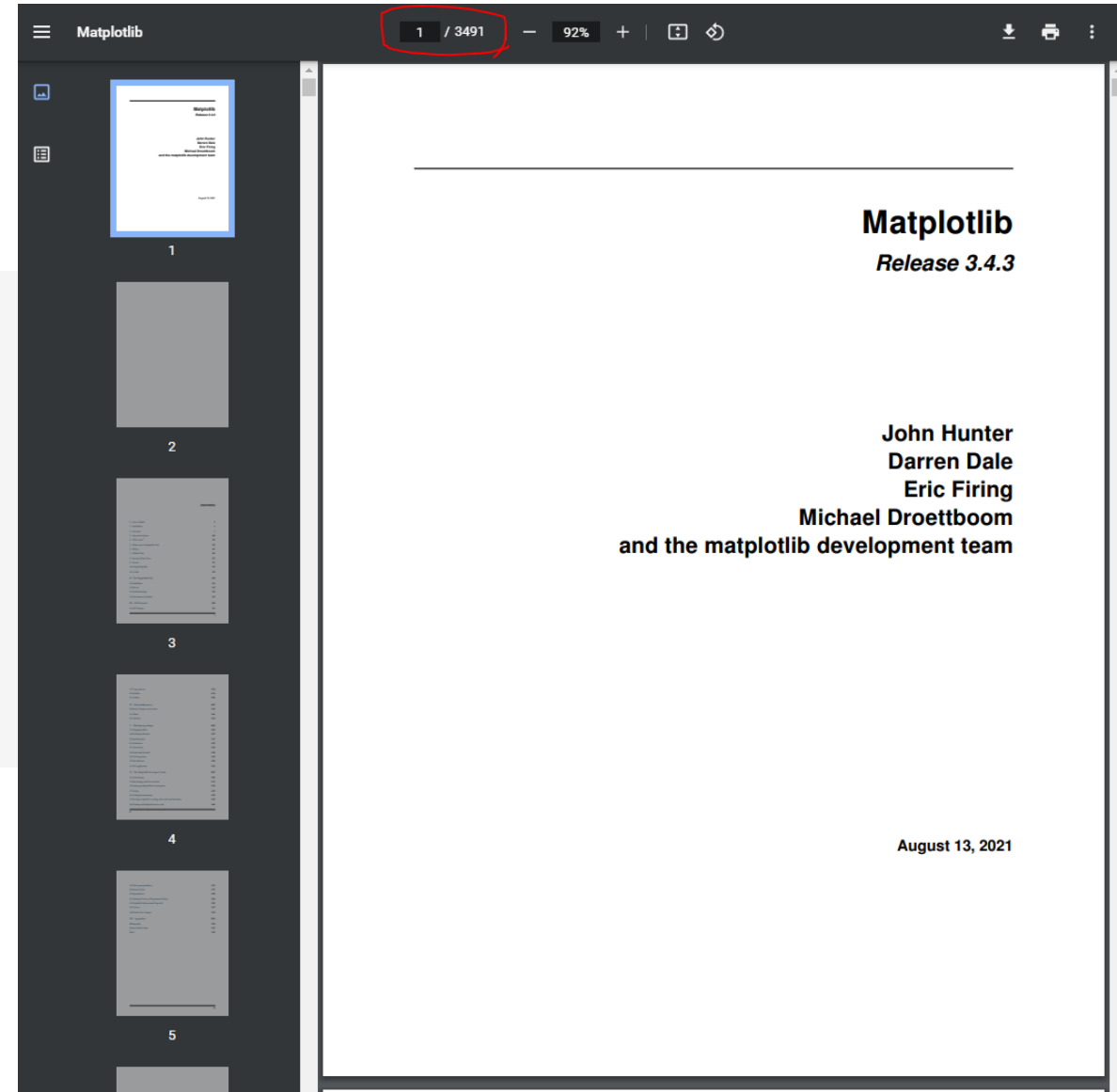
```
>>> import requests  
>>> import pandas as pd  
>>> import BeautifulSoup as bs
```

Python library

Python Standard Library



```
>>> from matplotlib import pyplot as plt  
>>> from bs4 import BeautifulSoup as bs
```

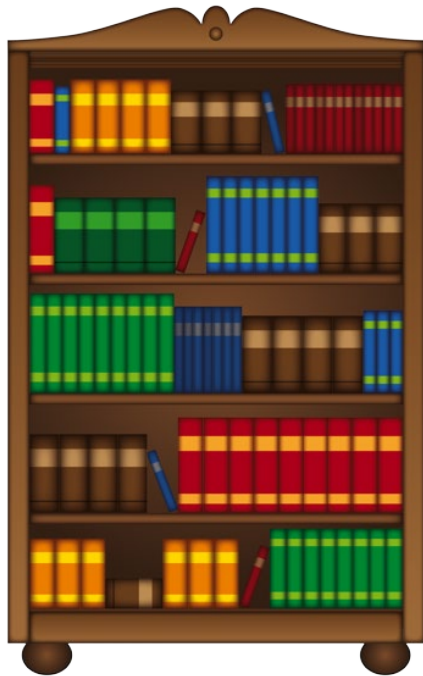


module not found error?!

```
import selenium
```

```
-----  
ModuleNotFoundError                                Traceback (most recent call last)  
<ipython-input-2-abb2a9e03f2a> in <module>  
----> 1 import selenium
```

```
ModuleNotFoundError: No module named 'selenium'
```



Need to stock your library
with a new book.



Anaconda Prompt (Anaconda3)

```
(base) C:\Users\hhsc8>pip install selenium  
Collecting selenium  
  Using cached https://files.pythonhosted.org/packages/4e/5d/d6/a7f5ca98459853/selenium-3.141.0-py2.py3-none-any.whl  
Requirement already satisfied: urllib3 in c:\users\hhsc8\appdata\local\anaconda\envs\base\lib\site-packages (1.24.2)  
Installing collected packages: selenium  
Successfully installed selenium-3.141.0  
  
(base) C:\Users\hhsc8>
```


for loop.

- Invitation letter to my friends.

```
>>> friends = ['lisa', 'jenny', 'alex', 'john']
```



defining a list.

```
>>>for friend in friends:
```

```
    print(friend)
```



defining a **for** loop:
tells Python to ..

- 1) get elements from the list of **friends**
- 2) and assign it to **friend**.



tell Python to print the name
which has been assigned to **friend**.

step 4.
Python repeats until the last element of the list.



a for loop inside a for loop.

```
>>> adj = ["red", "big", "tasty"]  
>>> fruits = ["apple", "banana", "cherry"]
```

```
>>> for x in adj:  
>>>     print(x)  
>>> for y in fruits:  
>>>     print(y)
```

```
red  
big  
tasty  
apple  
banana  
cherry
```

```
>>> for x in adj:  
>>>     for y in fruits:  
>>>         print(x,y)
```

```
red apple  
red banana  
red cherry  
big apple  
big banana  
big cherry  
tasty apple  
tasty banana  
tasty cherry
```



f-strings vs % formatting

- f-strings in Python 3.6: new way to format strings.
 - more readable, concise.

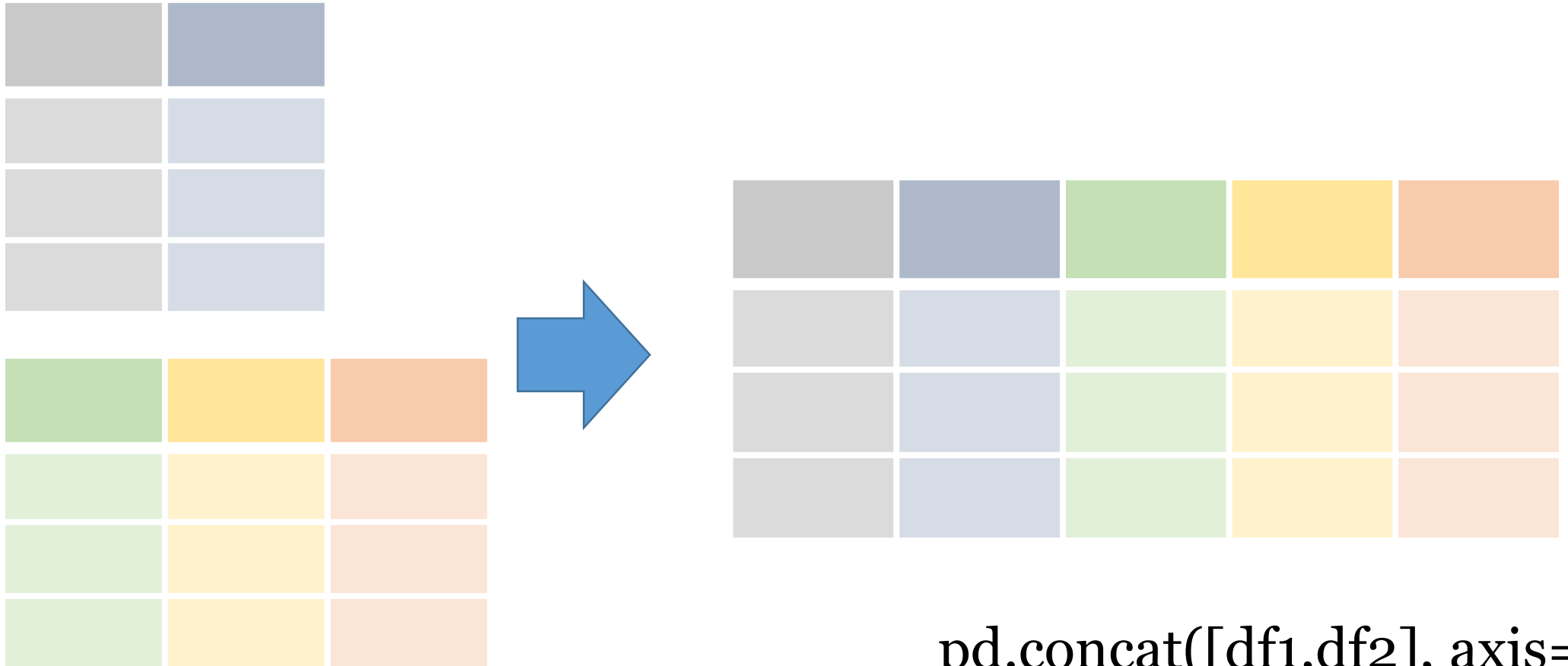
```
>>> name = "Heidi"
>>> coffee = "Americano"
>>> f"Hello, {name}, would you like to drink your {coffee}?"
'Hello Heidi, would you like to drink your Americano?'
```

```
>>> name = "Heidi"
>>> "Hello, %s." % name
Hello, Heidi
```



Concatenation function of pandas library

append columns of DataFrames



Encoding?!

- Basics: how computers store information
 - Computers use a binary system.
 - All data is represented in sequences of 1s and 0s.
 - Basic unit of binary is a bit. (single 1 or 0)
 - The largest unit of binary, a byte, consists of 8 bits.
 - Every digital stuff you see, from software to mobile apps, websites, Instagram etc is built on this system of bytes.
 - When we refer to file size, we are referencing the number of bytes.
 - 'TEXT' is represented in computers by a string of bits.



The American Standard Code for Information Interchange (ASCII)

- ASCII library include every upper and lower case letter in the Latin alphabet, every digit from 0-9 and some common symbols (/,.!?)
- There are 256 ways to group eight 1s and 0s together. (2^8)
- Was introduced in 1960. It was ok back then.
- But there are languages besides English.
- New systems were made to map other languages to the same set of 256 unique bytes.
- Having multiple encoding system is inefficient and confusing.
- So they developed 'Unicode' to store every symbol, and it is the universal standard for encoding all human languages and emojis.

ASCII characters with associated codes and bytes

Character	ASCII Code	BYTE
A	065	01000001
a	097	01100001
B	066	01000010
b	098	01100010
Z	090	01011010
z	122	01111010
0	048	00110000
9	057	00111001
!	033	00100001
?	063	00111111

Eg) The quick brown fox jumps over the lazy dog.

```
01010100 01101000 01100101 00100000 01110001 01110101
01101001 01100011 01101011 00100000 01100010 01110010
01101111 01110111 01101110 00100000 01100110 01101111
01111000 00100000 01101010 01110101 01101101 01110000
01110011 00100000 01101111 01110110 01100101 01110010
00100000 01110100 01101000 01100101 00100000 01101100
01100001 01111010 01111001 00100000 01100100 01101111
01100111 00101110
```

UTF-8 to the rescue

- UTF-8 is an encoding system for Unicode.
 - Unicode is an International Encoding Standard while UTF-8 is an encoding system.
- “UTF”, or “Unicode Transformation Format.”
- It can translate any Unicode character to a matching unique binary string, and can also translate the binary string back to a Unicode character.
- There are other encoding systems for Unicode besides UTF-8, but UTF-8 is unique because it represents characters in one-byte units.
- It is UTF “-8” because one byte consists of eight bits.
- UTF-8 is the most common character encoding method used on the internet today, and is the default character set for HTML5.
- Over 95% of all websites store characters this way.





Selenium

Limitations of BeautifulSoup

- BeautifulSoup has been a good web scraper starter for us.
- But! online retailers such as Amazon put anti-bot software so that you cannot crawl pages using BeautifulSoup. It will shut down requests coming from BeautifulSoup.
- Many websites will supply data that is dynamically loaded via javascript.
- So does that mean we have to do it manually?
- We can use...available Web Scraping tools.
- or.. we can try to automate our browsing behavior using selenium and chrome driver!



Selenium

- It was created primarily for automated web testing, but due to its compatibility with JavaScript, it is also used for web scraping.
- It is free to use and can be used across many platforms.
- Through Selenium...
 - your Google Chrome Browser mimics legitimate user browsing behaviors.
 - It will do the clicking and typing for you.
 - Scrape information from the websites
- But! It is slow.
- Selenium: complex projects
- BeautifulSoup: smaller projects



So let's go ahead and install....

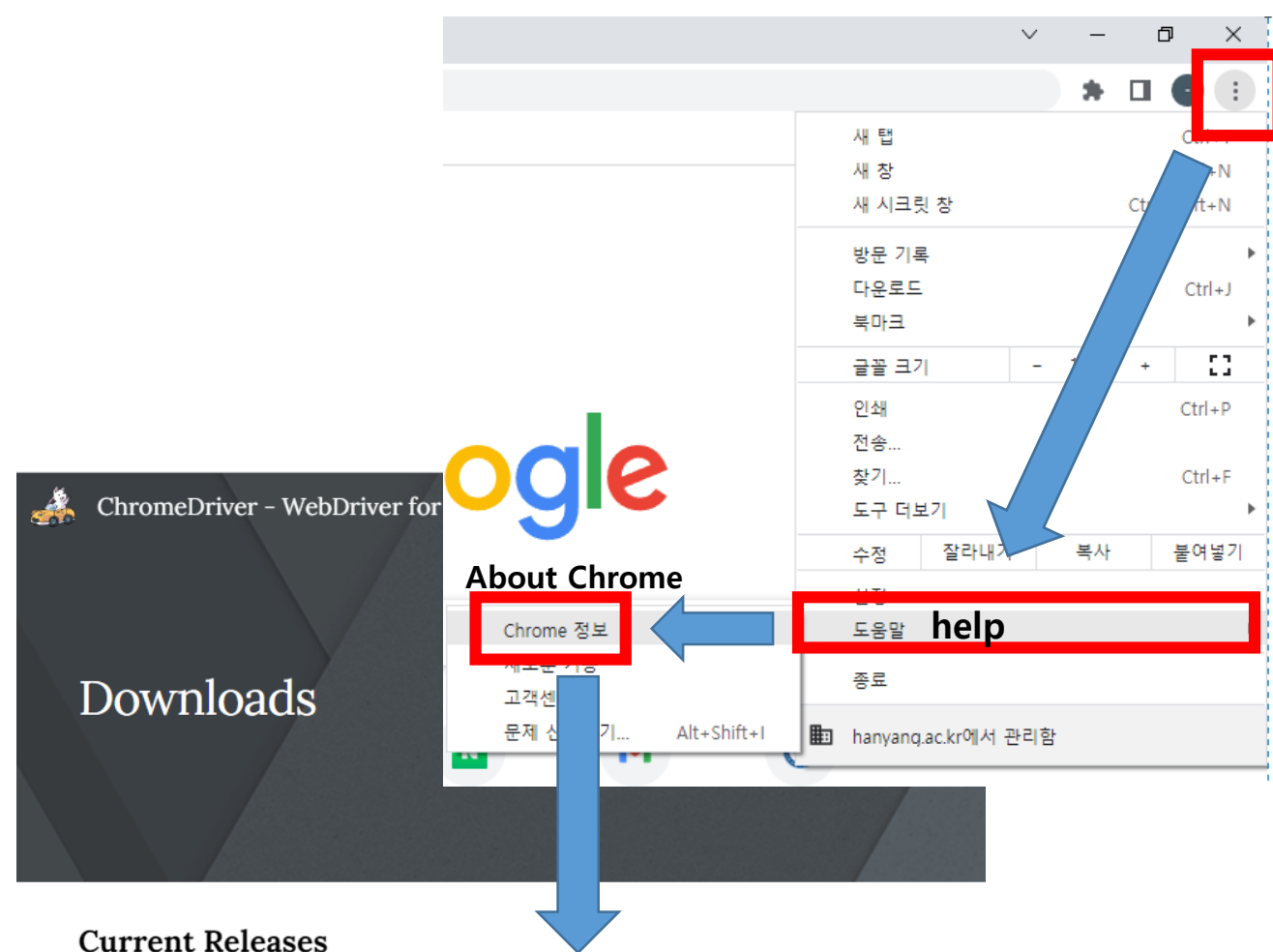
- ✓ ChromeDriver
- ✓ selenium



ChromeDriver

- Steps:

1. Check your chrome version
2. Download your ChromeDriver accordingly.
3. Unzip your ChromeDriver and place it in the same folder as your jupyter notebook file.



Current Releases

- If you are using Chrome version 106, please download [ChromeDriver 106.0.5249.21](#)
- If you are using Chrome version 105, please download [ChromeDriver 105.0.5195.52](#)
- If you are using Chrome version 104, please download [ChromeDriver 104.0.5112.79](#)
- For older version of Chrome, please see below for the version of ChromeDriver that supports it.

If you are using Chrome from Dev or Canary channel, please following instructions on the [ChromeDriver Canary](#) page.

For more information on selecting the right version of ChromeDriver, please see the [Version Selection](#) page.

selenium

- access through anaconda prompt

```
>>> pip list
```

- see whether you have selenium. If not,

```
>>> pip install selenium
```

- same goes for all the other libraries



Pygooglenews

- Google News RSS feed.
- We can get top stories, topic related news feeds, geolocation news feeds, and an extensive full text search feed.

The screenshot displays the Google News interface. At the top, there's a search bar with the text "Search for topics, locations & sources". Below this, a banner states "The new Google News has a fresh look, brand-new briefing, & customized topics. Try it out". The left sidebar contains navigation options: "Top stories", "For you", "Following", "Saved searches", "COVID-19", "U.S.", "World", "Your local news", "Business", "Technology", "Entertainment", "Sports", "Science", and "Health". The main content area features "Headlines" with several news items, including "COVID-19 news: See the latest coverage of the coronavirus", "Four takeaways from New Hampshire and Rhode Island primaries", "Northeastern University explosion: Package explodes on Boston campus; 1 injured, FBI involved", and "Queen Elizabeth's coffin was made 30 years ago with lead, English oak". On the right, there's a "Your local weather" section showing "Partly cloudy 21°C" and a "Fact check" section with items like "Challenger Barnes vastly overstates case, suggesting U.S. Sen. Johnson backs outright abortion ban" and "Fact Check: Is Donald Trump Not Invited to Queen Elizabeth II's Funeral?".

Google News

Search for topics, locations & sources

The new Google News has a fresh look, brand-new briefing, & customized topics. [Try it out](#)

Top stories

- For you
- Following
- Saved searches
- COVID-19
- U.S.
- World
- Your local news
- Business
- Technology
- Entertainment
- Sports
- Science
- Health

Language & region
English (United States)

Settings

Get the Android app

Get the iOS app

Send feedback

Help

Headlines [More Headlines](#)

COVID-19 news: See the latest coverage of the coronavirus

Four takeaways from New Hampshire and Rhode Island primaries
CNN · 3 hours ago

- New Hampshire, Rhode Island primary election results and news for 2022 midterms
Fox News · 11 hours ago
- What's at stake for Republicans, Democrats in midterms now that primary election season is over
Fox News · 20 hours ago · Opinion
- New Hampshire holds primary Tuesday as Pennsylvania Senate campaign heats up
CBS News · 10 hours ago

[View Full Coverage](#)

Northeastern University explosion: Package explodes on Boston campus; 1 injured, FBI involved
WPVI-TV · 37 minutes ago

- FBI investigates explosion at Northeastern University
CBS Boston · 7 hours ago

[View Full Coverage](#)

Queen Elizabeth's coffin was made 30 years ago with lead, English oak
USA TODAY · 17 hours ago

Your local weather

Partly cloudy
21°C

Today Thu Fri Sat Sun

25°C 27°C 28°C 28°C 27°C
18°C 17°C 18°C 18°C 19°C

C | F | K [More on weather.com](#)

Fact check

Challenger Barnes vastly overstates case, suggesting U.S. Sen. Johnson backs outright abortion ban
PolitiFact · 19 hours ago

Fact Check: Is Donald Trump Not Invited to Queen Elizabeth II's Funeral?
Newsweek · 19 hours ago

Posts spread false claims about Queen Elizabeth II's corgis
AFP Factcheck · 14 hours ago

False: Russia is Merely 'Regrouping' After its Rout in Kharkiv
Polygraph.info · 18 hours ago

HTML vs XML

HTML	XML
HTML stands for Hyper Text Markup Language.	XML stands for extensible Markup Language.
Focusses on the appearance of data . Enhances the display of text.	The main purpose is to focus on the transport of data and saving the data .
HTML is static in nature.	XML is dynamic in nature.
HTML is a markup language.	XML provides framework to define markup languages.
HTML is not Case sensitive.	XML is Case sensitive.
HTML tags are predefined tags. (<title>,<head>,<body>)	XML tags are user defined tags.
There are limited number of tags in HTML.	XML tags are extensible.
HTML does not preserve white spaces.	White space can be preserved in XML.
HTML tags are used for displaying the data.	XML tags are used for describing the data not for displaying.
HTML is used to display the data. HTML is presentation driven. How the text appears is of utmost importance.	XML is used to store data. XML is content-driven and not many formatting features are available.
HTML does not carry data it just display it.	XML carries the data to and from database.
HTML document size is relatively small.	XML document size is relatively large as the approach of formatting and the codes both are lengthy.



XPath

- XPath stands for XML Path Language, and the most important function of XPath is that **it tells us the path, or address, of the various pieces of data in the XML file.**
- XPath is a syntax for defining parts of an XML document.
- XPath uses path expressions to navigate in XML documents.
- XPath contains a library of standard functions.
- All of the information that is transferred between the web server and the computers we use can be saved as **XML** files, and we can then use the **XPath** query language to search the XML data files, and even compute values based on selected criteria.





Let's launch our jupyter notebook and try it for ourselves

```
>>> pip install selenium
```

```
>>> pip install pygooglenews
```

Other site that you can play around with

- <http://books.toscrape.com/>

