데이터융합SW과
김규석 교수

Data Analysis with Java

데이터 분석 프로그래밍02

Objective of Today's Class

Data Crawling

- Collect unstructured data with Selenium

Collect Unstructured Data(Cont’d)

Selenium

- A portable framework for testing web applications

Source：https://en.wikipedia.org/wiki/Selenium_(software) / https://www.selenium.dev/

Collect Unstructured Data(Cont'd)

Download a Selenium client and add it to the project

- https://www.selenium.dev/downloads/

Selenium Client & WebDriver Language Bindings

In order to create scripts that interact with the Selenium Server (Remote WebDriver) or create local Selenium WebDriver scripts, you need to make use of language-specific client drivers.

While language bindings for **other languages exist**, these are the core ones that are supported by the main project hosted on GitHub.

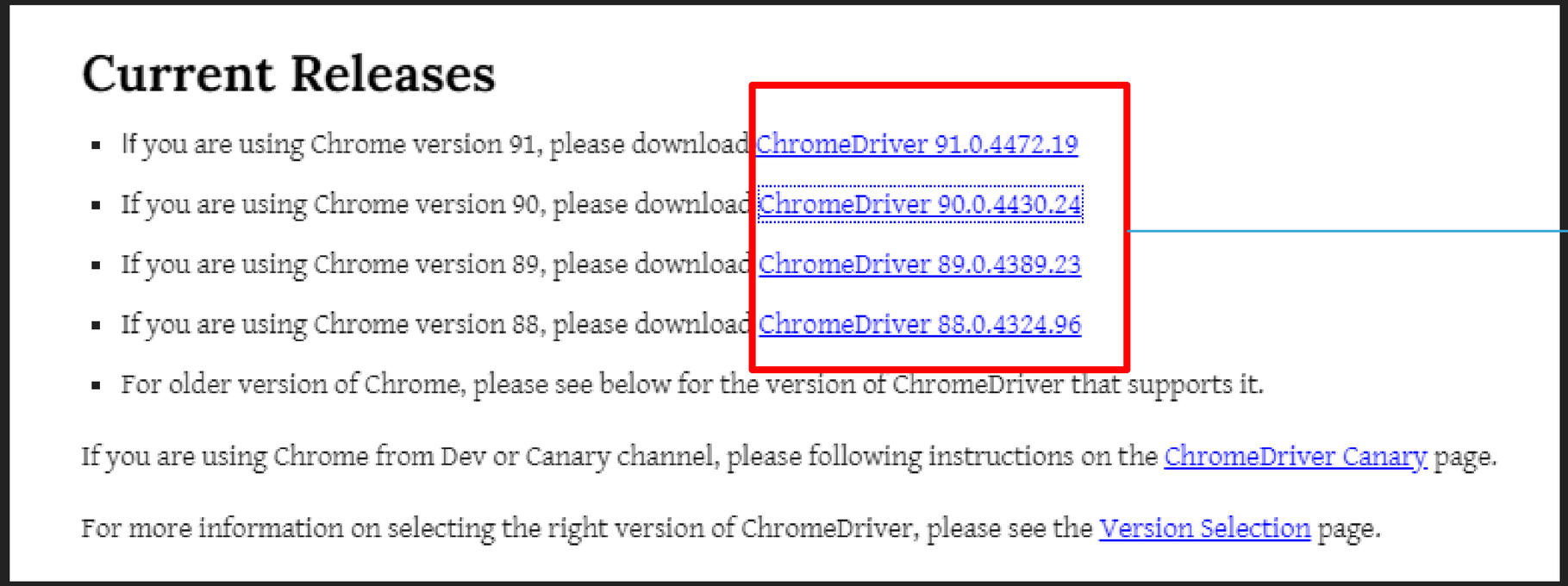
LANGUAGE	STABLE VERSION	RELEASE DATE	BETA VERSION	BETA RELEASE DATE	LINKS
Ruby	3.142.6	October 04, 2019	4.0.0beta3	April 13, 2021	Download Beta Download Changelog API Docs
Java	3.141.59	November 14, 2018	4.0.0-beta-3	April 13, 2021	Download Beta Download Changelog API Docs
Python	3.141.0	November 01, 2018	4.0.0.b3	April 13, 2021	Download Beta Download Changelog API Docs
C#	3.14.0	August 02, 2018	4.0.0-beta2	March 17, 2021	Download Beta Download Changelog API Docs
JavaScript	3.6.0	October 06, 2017	4.0.0-beta.3	April 13, 2021	Download Beta Download Changelog API Docs

Source：https://en.wikipedia.org/wiki/Selenium_(software) / https://www.selenium.dev/

Collect Unstructured Data(Cont'd)

Download a ChromeDriver

- https://chromedriver.chromium.org/downloads



The versions of the ChromeDriver and Chrome brower should be the same

Collect Unstructured Data(Cont'd)

Open a URL and Get the Page Source

```
6  public class Main {
7      public static WebDriver driver;
8      public static String base_url = "https://www.naver.com";
9      public static final String WEB_DRIVER_ID = "webdriver.chrome.driver";
10     public static final String WEB_DRIVER_PATH = "C:/Users/CTC/Downloads/chromedriver_win32/chromedriver.exe";
11
12     public static void main(String[] args) {
13         System.setProperty(WEB_DRIVER_ID, WEB_DRIVER_PATH);
14         driver = new ChromeDriver();
15         crawl();
16     }
17
18     public static void crawl() {
19         try {
20             driver.get(base_url);
21             System.out.println(driver.getPageSource());
22         } catch (Exception e) {
23             e.printStackTrace();
24         } finally {
25             driver.close();
26         }
27     }
28 }
```

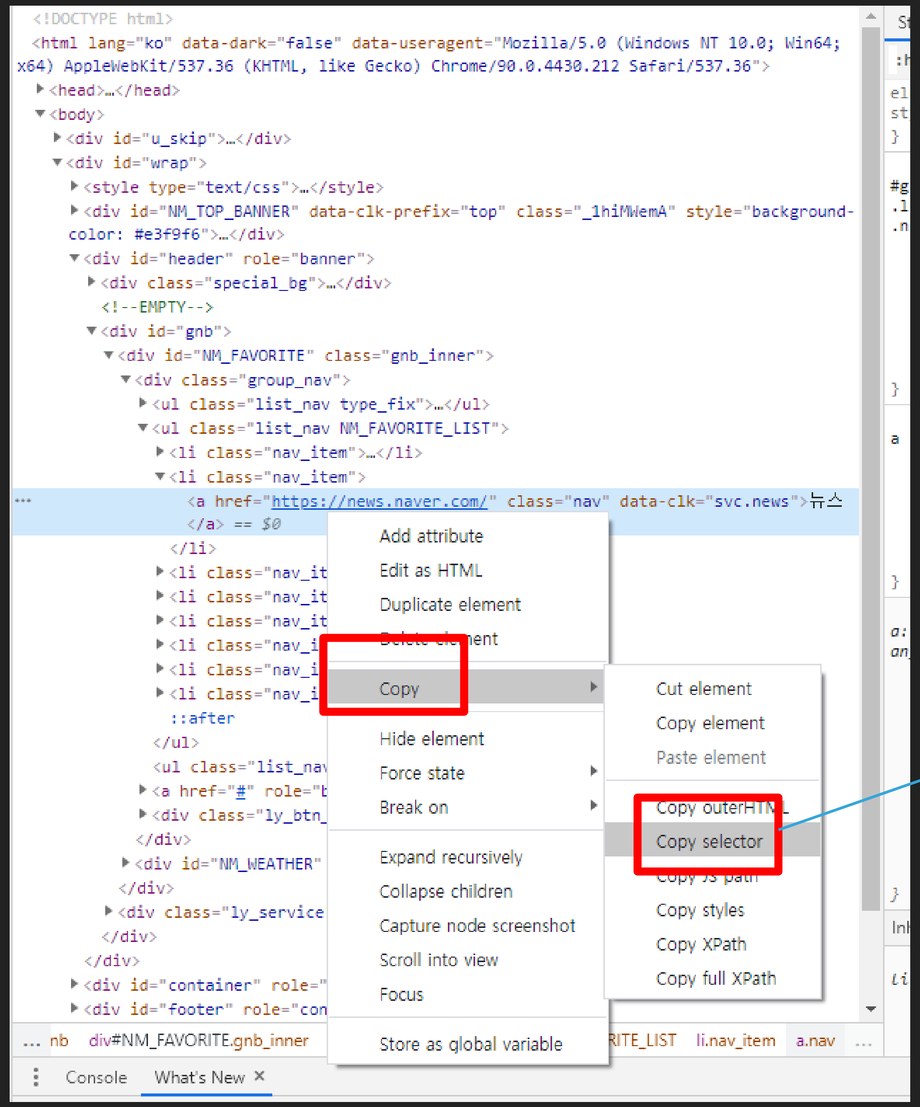
Collect Unstructured Data(Cont'd)

Click on a Button

Source : https://www.naver.com/

Collect Unstructured Data(Cont'd)

Click on a Button

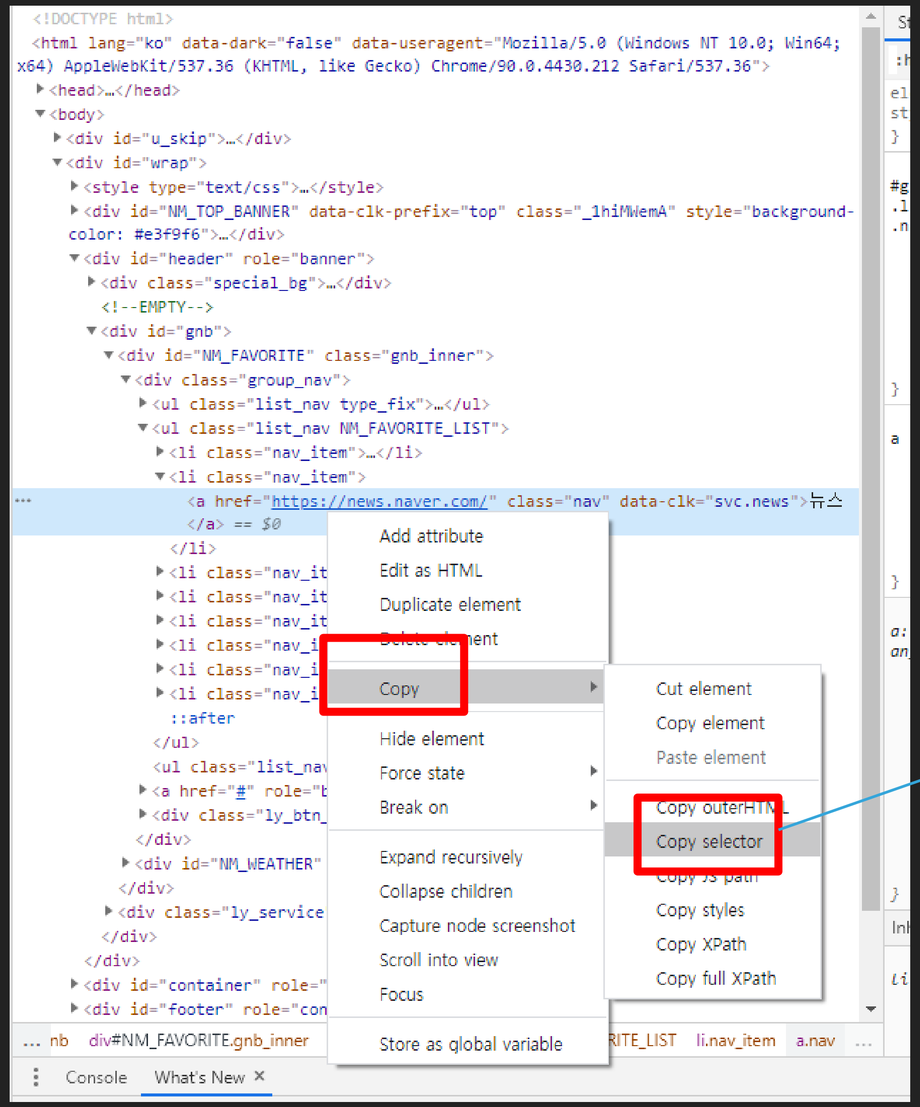


#NM_FAVORITE > div.group_nav > ul.list_nav.NM_FAVORITE_LIST > li:nth-child(2) > a

Source：https://www.naver.com/

Collect Unstructured Data

Click on a Button




```
public static void crawl() {
    try {
        driver.get(base_url);
        Thread.sleep(3000);
        driver.findElement(By.cssSelector("#NM_FAVORITE > div.group_nav > ul.list_nav.NM_FAVORITE_LIST > li:nth-child(2) > a")).click();
        Thread.sleep(3000);
    } catch (Exception e) {
        e.printStackTrace();
    } finally {
        driver.close();
    }
}
```

#NM_FAVORITE > div.group_nav > ul.list_nav.NM_FAVORITE_LIST > li:nth-child(2) > a

Source : https://www.naver.com/

P1

Collect Unstructured Data

- Crawl the text below

P2

Collect Titles of My Favorite Idol News Articles

- Search for the news titles which were published for the last 1 week

- Restore the data into a CSV file (No., Title, Date)

P3

Compose an Automated Testing Tool

- Target site : a bulletin board which was designed and composed in class

- Function
  - Click on the “write” button
  - Insert random values into the fields
  - Click on the “submit” button
  - The processes above should be done 1,000 times

Keywords

What we need to understand are ...

- Correlation coefficient r
- R^2
- Adj. R^2
- Coefficient
- Significance Level