

7주차 R 프로그래밍

2019. 2. 27

Geon

✓ 목차

1. 웹 크롤러 실습2
2. 데이터 저장 및 분석

2. 웹 크롤링 실습

rvest

실습1. 네이버 뉴스 목록 크롤링



'기어S3 LTE' 스마트폰 없이 단독 결제

전국 오프라인 매장에서 사용 가능한 스마트워치 단독 결제 서비스가 국내 처음 개시 ...
전자신문 | 1시간전



“PC와 출입증이 사라졌다”..SK텔레콤, ‘5G스마트오피스’ 첫 적용

[이데일리 김현아 기자] SK텔레콤의 MKT Data사업팀 등 300여 명이 일하...

이데일리 | 1시간전



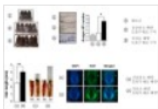
방통위, 통신장애 배상·음란물 유통 '규제강화'

[아이뉴스24 김문기 기자] 방송통신위원회가 통신장애 등에 따른 손해배상 규정을 ...
아이뉴스24 | 1시간전



삼성전자, 전 세계에 '갤럭시 언팩' 한글 광고한 이유는?

- 화제성 극대화...글로벌 주요 랜드마크서 진행 [디지털데일리 윤상호기자] 삼성전자 ...



모발 생성 세포 증식 탈모 세포치료제 길터...2020년 임상시험

모발 생성에 관여하는 세포를 이용한 세포치료제 개발에 청신호가 켜졌다. 국내 연구 ...
조선비즈 | 1시간전

① AiRS 추천으로 구성된 뉴스를 제공합니다.

1 2 3 4 5 6 7 8 9 10 | 다음>

```
<td class="content">  
  <div id="main_content" class="content">  
    <div class="list_body section_index">  
      <div class="cluster">...</div>  
      <div class="_persist">...</div>  
      <div class="cluster">...</div>  
      <div class="cluster">...</div>  
      <div class="section_issue_list">...</div>  
      <div class="section_ad">...</div>  
      <a id="mainNewsComponentId"></a>  
      <h4 class="blind">뉴스 더보기</h4>  
    <div class="section_body" id="section_body">  
      <ul class="type06_headline">  
        <li>  
          <dl>  
            <dt class="photo">...</dt>  
            <dt>  
              <a href="/main/read.nhn?mode=LSD&mid=shm&sid1=10  
                'nclicks(itn.airscnt,'880000EA_000000000000000000  
                'BDrMmEUa7btfx5Zd')">화웨이, 올해 폴더블폰 20만대  
            </dt>  
            <dd class data-comment="{gno:'news030,0002784097',  
              'sectionHomeList'}">...</dd>  
          </dl>  
        </li>
```

2. 웹 크롤링 실습

rvest

실습1. 네이버 뉴스 목록 크롤링

1. `library(rvest)`
2. `library(dplyr)`
3. `## 뉴스 목록 크롤링`
4. `newsURL <- "https://news.naver.com/main/main.nhn?mode=LSD&mid=shm&sid1=105"`
5. `newsHtml <- read_html(newsURL)`
6. `newsPgList <- newsHtml %>%`
7. `html_nodes("#main_content") %>%`
8. `html_nodes("#section_body") %>%`
9. `html_node("ul")`

```
> newsPgList  
{xml_nodeset (1)}  
[1] <NA>
```

텍스트 데이터 없음

1. 웹 크롤링 실습2

RSelenium

실습2. Selenium 설치

selenium-server-standalone-3.141.59.jar 다운로드

<https://www.seleniumhq.org/>



The screenshot shows the SeleniumHQ website. At the top, there is a navigation bar with links: Projects, Download, Documentation, Support, and About. The 'Download' link is highlighted with a red box. Below the navigation bar, the page title is 'SeleniumHQ Browser Automation'. The main content area is divided into two columns. The left column is titled 'What is Selenium?' and contains text about Selenium's purpose and its support from major browser vendors. The right column is titled 'Which part of Selenium is appropriate for me?' and contains two sections: 'Selenium WebDriver' and 'Selenium IDE'. Each section has a list of bullet points describing its capabilities. At the bottom right, there is a 'Download Selenium' button and a 'Donate to Selenium' section with a 'Donate' button and a list of sponsors.

SeleniumHQ
Browser Automation

edit this page | search selenium: Go

Projects Download Documentation Support About

What is Selenium?

Selenium automates browsers. That's it! What you do with that power is entirely up to you. Primarily, it is for automating web applications for testing purposes, but is certainly not limited to just that. Boring web-based administration tasks can (and should!) be automated as well.

Selenium has the support of some of the largest browser vendors who have taken (or are taking) steps to make Selenium a native part of their browser. It is also the core technology in countless other browser automation tools, APIs and frameworks.

Which part of Selenium is appropriate for me?

Selenium WebDriver



If you want to

- create robust, browser-based regression automation suites and tests
- scale and distribute scripts across many environments

Then you want to use [Selenium WebDriver](#); a collection of language specific bindings to drive a browser -- the way it is meant to be driven.

Selenium WebDriver is the successor of [Selenium Remote Control](#) which has been officially deprecated. The Selenium Server (used by both WebDriver and Remote Control) now also includes built-in grid capabilities.

Selenium IDE



If you want to

- create quick bug reproduction scripts
- create scripts to aid in automation-aided exploratory testing

Then you want to use [Selenium IDE](#); a Chrome and Firefox add-on that will do simple record-and-playback of interactions with the browser.



Selenium is a suite of tools to automate web browsers across many platforms.

Selenium...

- runs in [many browsers](#) and [operating systems](#)
- can be controlled by many [programming languages](#) and [testing frameworks](#).

[Download Selenium](#)

Donate to Selenium

with PayPal

[Donate](#)

VISA MASTERCARD AMERICAN EXPRESS DISCOVER

through sponsorship

You can [sponsor the Selenium project](#) if you'd like some public recognition of your generous contribution.

Selenium Sponsors

Want to support the Selenium project? [Learn more about sponsorship](#) or view the [full list of sponsors](#).

1. 웹 크롤링 실습2

RSelenium

실습2. Selenium 설치

selenium-server-standalone-3.141.59.jar 다운로드

<https://www.seleniumhq.org/>

SeleniumHQ
Browser Automation

edit this page search selenium: [] Go

Projects Download Documentation Support About

Downloads

Below is where you can find the latest releases of all the Selenium components. You can also find a list of [previous releases](#), [source code](#), and additional information for [Maven users](#) (Maven is a popular Java build tool).

Selenium Standalone Server

The Selenium Server is needed in order to run Remote Selenium WebDriver. Selenium 3.X is no longer capable of running Selenium RC directly, rather it does it through emulation and the [WebDriver backed Selenium interface](#).

Download version **3.141.59**

To run Selenium tests exported from the legacy IDE, use the [Selenium Html Runner](#).

To use the Selenium Server in a Grid configuration [see the wiki page](#).

The Internet Explorer Driver Server

This is required if you want to make use of the latest and greatest features of the WebDriver InternetExplorerDriver. Please make sure that this is available on your \$PATH (or %PATH% on Windows) in order for the IE Driver to work as expected.

Download version 3.14.0 (for recommended) [32 bit Windows IE](#) or [64 bit Windows IE](#) [CHANGELOG](#)

Selenium Client & WebDriver Language Bindings

In order to create scripts that interact with the Selenium Server (Selenium RC, Selenium Remote WebDriver) or create local Selenium WebDriver scripts, you need to make use of language-specific client drivers. These languages include both 1.x and 2.x style clients.

While language bindings for [other languages exist](#), these are the core ones that are supported by the main project hosted on GitHub.

Language	Client Version	Release Date	Download	Change log	Javadoc
Java	3.141.59	2018-11-14	Download	Change log	Javadoc
C#	3.14.0	2018-08-02	Download	Change log	API docs
Ruby	3.14.0	2018-08-03	Download	Change log	API docs
Python	3.14.0	2018-08-02	Download	Change log	API docs
Javascript (Node)	4.0.0-alpha.1	2018-01-13	Download	Change log	API docs

C# NuGet

NuGet latest release is 3.14.0, Released on 2018-08-02

- [WebDriver](#)
- [WebDriverBackedSelenium](#)
- [Support](#)
- [RC](#) (Final version 3.1.0 Released 2017-02-16)

SafariDriver - DEPRECATED - use Apple's SafariDriver with Safari 10+

SafariDriver now requires manual installation of the extension prior to automation

Selenium Downloads

[Previous Releases](#)
[Source Code](#)
[Maven Information](#)

Donate to Selenium

with PayPal

Donate

VISA

through sponsorship

You can [sponsor the Selenium project](#) if you'd like some public recognition of your generous contribution.

Selenium Sponsors

See who [supports the Selenium project](#).

BrowserStack

SAUCE LABS

experitest
Selenium for Mobile

New Relic
SYNTHETICS
WEBSITE MONITORING WITH SELENIUM

CrossBrowserTesting
A SMARTBEAR COMPANY

applitools
Automated Visual Testing

1. 웹 크롤링 실습2

RSelenium

실습2. Selenium 설치

geckodriver.exe 다운로드

<https://github.com/mozilla/geckodriver>

mozilla / geckodriver

Watch 311 Star 2,775 Fork 592

Code Issues 316 Pull requests 0 Insights

WebDriver for Firefox <https://firefox-source-docs.mozilla.org/>

webdriver geckodriver firefox rust gecko

391 commits 3 branches 31 releases 17 contributors

Branch: master New pull request Create new file Upload files Find file Clone or download

andreastt import 4d6d3403eb6b015ebd2e6949d57dd518d07d024f Latest commit fa75e69 on 16 Nov 2018

CONTRIBUTING.md	import of ba6208ac98c6bc52fab16237571a95d64be64755	5 months ago
ISSUE_TEMPLATE.md	Update issue template to request trace logs as attachment	5 months ago
README.md	import 4d6d3403eb6b015ebd2e6949d57dd518d07d024f	3 months ago

README.md

geckodriver

Proxy for using W3C [WebDriver](#) compatible clients to interact with Gecko-based browsers.

This program provides the HTTP API described by the [WebDriver protocol](#) to communicate with Gecko browsers, such as Firefox. It translates calls into the [Firefox remote protocol](#) by acting as a proxy between the local- and remote ends.

geckodriver's [source code](#) is made available under the [Mozilla Public License](#).

Downloads

- [Releases](#)
- [Change log](#)

스크롤 내리기

1. 웹 크롤링 실습2

RSelenium

실습2. Selenium 설치

geckodriver.exe 다운로드

<https://github.com/mozilla/geckodriver>

This allows WebDriver to be used with various popular web frameworks that—through indirection—hides the file upload control and invokes it through other means.

- Allow use of an indefinite script timeout for the [Set Timeouts](#) command, thanks to reimu.

Fixed

- Corrected `Content-Type` of response header to `utf-8` to fix an HTTP/1.1 compatibility bug.
- Relaxed the deserialization of timeouts parameters to allow unknown fields for the [Set Timeouts](#) command.
- Fixed a regression in the [Take Element Screenshot](#) to not screenshot the viewport, but the requested element.

Assets 7

 geckodriver-v0.24.0-linux32.tar.gz	2.78 MB
 geckodriver-v0.24.0-linux64.tar.gz	2.76 MB
 geckodriver-v0.24.0-macos.tar.gz	2 MB
 geckodriver-v0.24.0-win32.zip	3.73 MB
 geckodriver-v0.24.0-win64.zip	4.46 MB
 Source code (zip)	
 Source code (tar.gz)	

1. 웹 크롤링 실습2

RSelenium

실습2. Selenium 설치

chromedriver.exe 다운로드

<http://chromedriver.chromium.org/downloads>

The screenshot shows the official ChromeDriver website. On the left is a navigation menu with links like CHROMEDRIVER, CAPABILITIES & CHROMEOptions, CHROME EXTENSIONS, CHROMEDRIVER CANARY, CONTRIBUTING, DOWNLOADS, GETTING STARTED, LOGGING, MOBILE EMULATION, and NEED HELP?. The main content area is titled 'Downloads' and features a section for 'Current Releases' which is highlighted with a red box. This section contains a list of instructions for downloading the correct version of ChromeDriver based on the Chrome version being used. To the right of the 'Current Releases' section, there is a blue text annotation that reads '크롬 버전에 따라 다르게 다운로드' (Download differently according to Chrome version).

ChromeDriver - WebDriver for Chrome

Search this site

Downloads

Current Releases 크롬 버전에 따라 다르게 다운로드

- If you are using Chrome version 73, please download [ChromeDriver 73.0.3683.20](#)
- If you are using Chrome version 72, please download [ChromeDriver 2.46](#) or [ChromeDriver 72.0.3626.69](#)
- If you are using Chrome version 71, please download [ChromeDriver 2.46](#) or [ChromeDriver 71.0.3578.137](#)
- 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 download [ChromeDriver 2.46](#). This is not officially supported, but in most cases it should work without major issues.

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

1. 웹 크롤링 실습2






RSelenium

실습2. Selenium 설치

chromedriver.exe 다운로드

<http://chromedriver.chromium.org/downloads>

Index of /2.46/

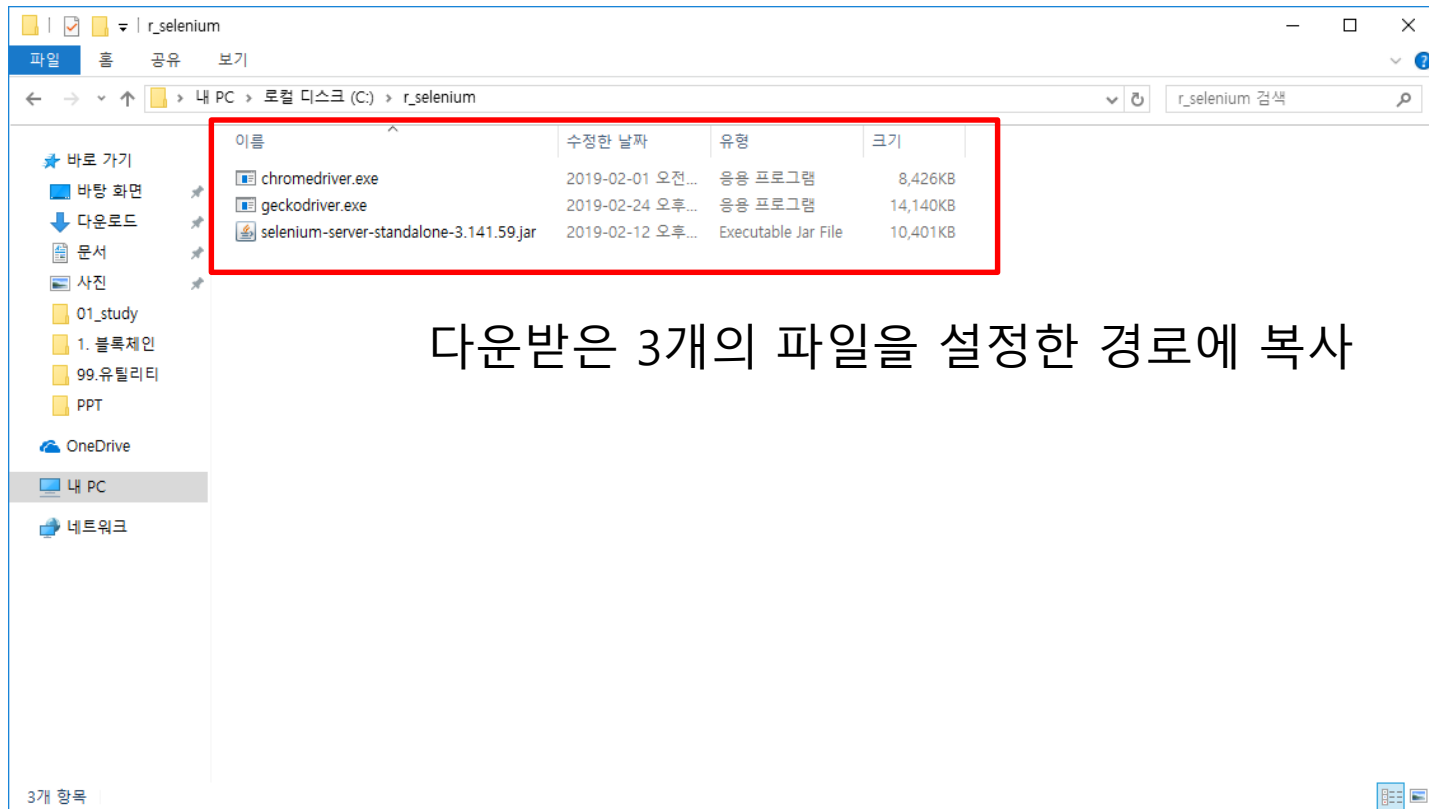
	<u>Name</u>	Last modified	Size	ETag
	Parent Directory		-	
	chromedriver_linux64.zip	2019-02-01 19:22:24	5.15MB	f63b50301dbce2335cdd442642d7efa0
	chromedriver_mac64.zip	2019-02-01 21:35:33	6.73MB	e287c1b628fbd9f6092ddd0353cbdda f
	chromedriver_win32.zip	2019-02-01 21:20:53	4.42MB	d498f2bb7a14216b235f122a615df07a
	notes.txt	2019-02-01 21:41:08	0.02MB	3cee5a7e5102a1fe996a7bb84c52983f

1. 웹 크롤링 실습2

RSelenium

실습2. Selenium 설정

C:\Wr_selenium (파일 경로는 자유롭게 진행 가능)

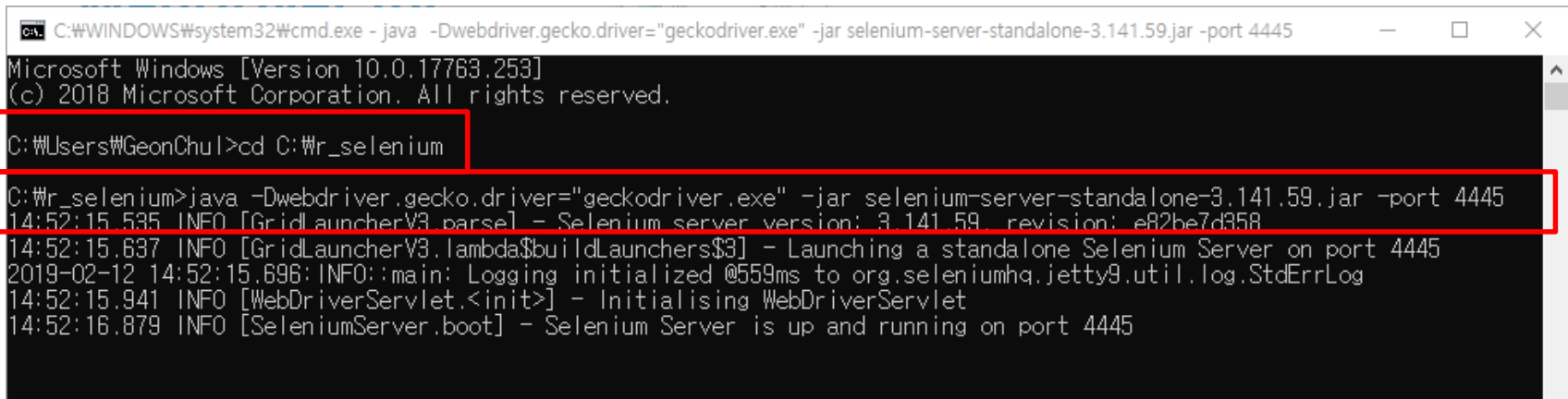


1. 웹 크롤링 실습2

RSelenium

실습2. geckodriver 실행

- ① cmd 실행
- ② cd C:\Wr_selenium
- ③ java -Dwebdriver.gecko.driver="geckodriver.exe" -jar selenium-server-standalone-3.141.59.jar -port 4445



```
C:\WINDOWS\system32\cmd.exe - java -Dwebdriver.gecko.driver="geckodriver.exe" -jar selenium-server-standalone-3.141.59.jar -port 4445
Microsoft Windows [Version 10.0.17763.253]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\GeonChul>cd C:\Wr_selenium

C:\Wr_selenium>java -Dwebdriver.gecko.driver="geckodriver.exe" -jar selenium-server-standalone-3.141.59.jar -port 4445
14:52:15.535 INFO [GridLauncherV3.parse] - Selenium server version: 3.141.59, revision: e82be7d358
14:52:15.637 INFO [GridLauncherV3.lambda$buildLaunchers$3] - Launching a standalone Selenium Server on port 4445
2019-02-12 14:52:15.696: INFO::main: Logging initialized @559ms to org.seleniumhq.jetty9.util.log.StdErrLog
14:52:15.941 INFO [WebDriverServlet.<init>] - Initialising WebDriverServlet
14:52:16.879 INFO [SeleniumServer.boot] - Selenium Server is up and running on port 4445
```

1. 웹 크롤링 실습2

RSelenium

실습3. RSelenium을 활용한 크롤링 만들기

- ① library(RSelenium)
- ② library(rvest)
- ③ remoteDr <- remoteDriver(remoteServerAddr = "localhost",
port = 4445L, browserName = "chrome")
- ④ remoteDr\$open()



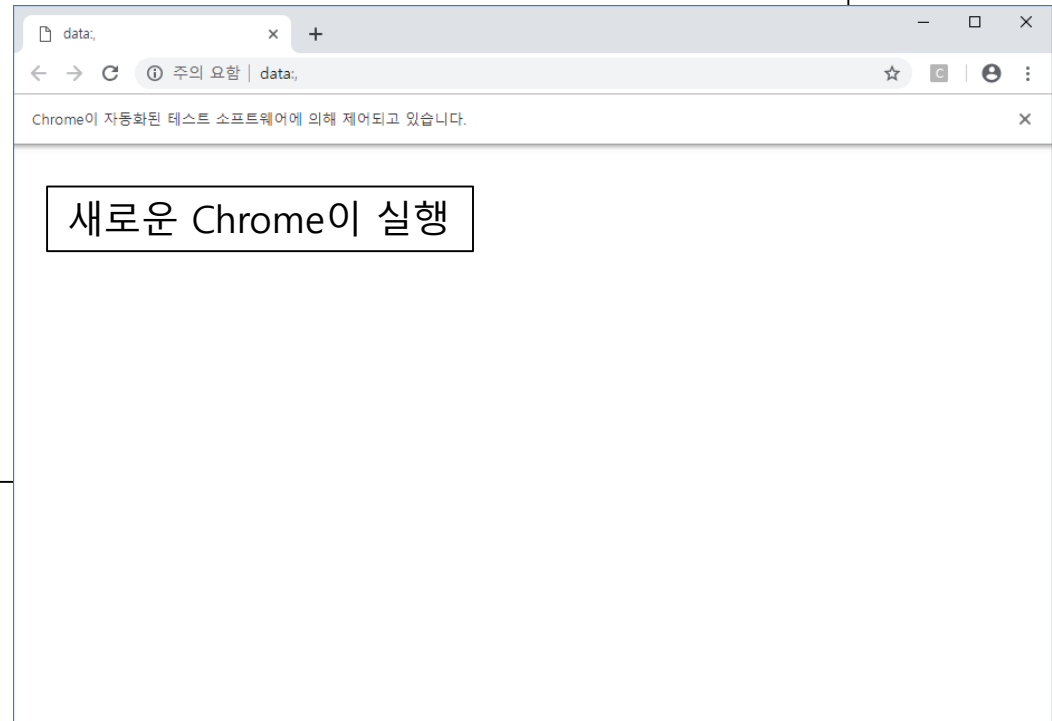
1. 웹 크롤링 실습2

RSelenium

실습3. RSelenium을 활용한 크롤링 만들기

```
> remoteDr <- remoteDriver(remoteServerAddr = "localhost", port = 4445L, browserName =  
  "chrome")  
> remoteDr$open()  
[1] "Connecting to remote server"  
$acceptInsecureCerts  
[1] FALSE  
  
$acceptSslCerts  
[1] FALSE  
  
$applicationCacheEnabled  
[1] FALSE  
  
$browserConnectionEnabled  
[1] FALSE  
  
$browserName  
[1] "chrome"  
  
$chrome
```

실행결과

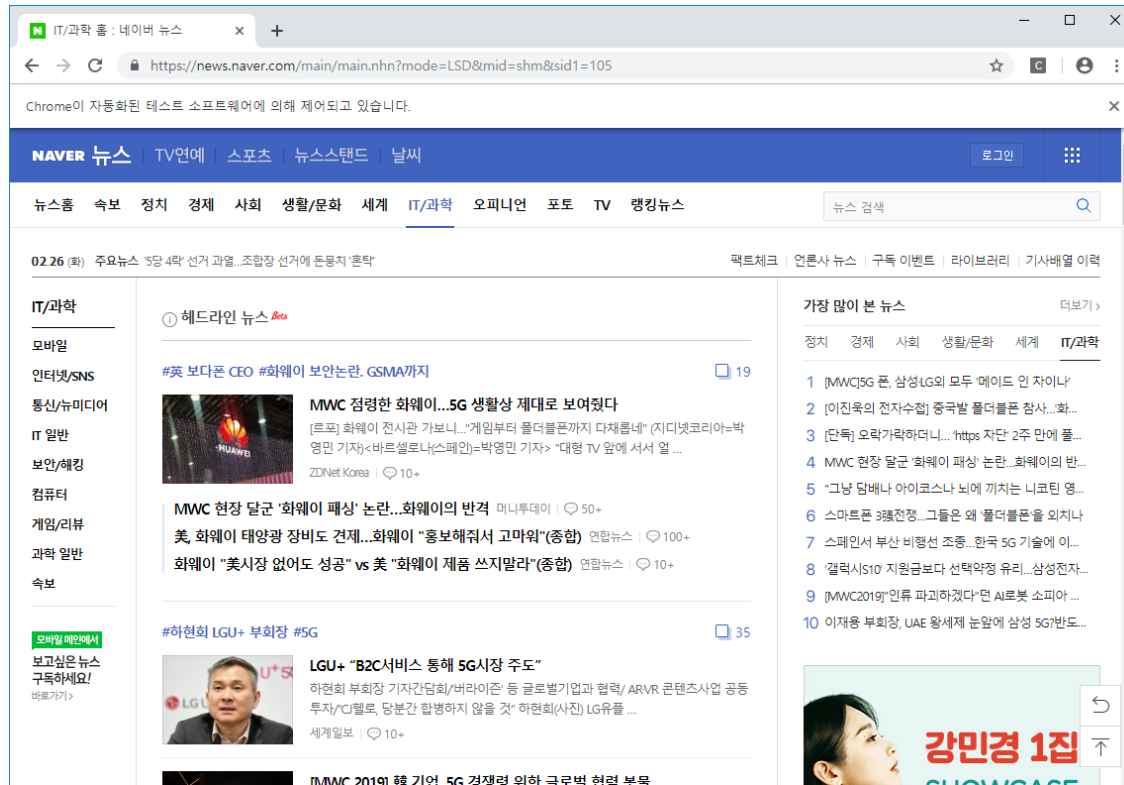


1. 웹 크롤링 실습2

R Selenium

실습3. RSelenium을 활용한 크롤링 만들기

⑤ `remoteDr$navigate("https://news.naver.com/main/main.nhn?mode=LSD&mid=shm&sid1=105")`



자동으로 URL
이동되는 것을 확인 가능

1. 웹 크롤링 실습2

R Selenium

실습3. RSelenium을 활용한 크롤링 만들기

- ⑥ `htmlFull <- remoteDr$getPageSource()[[1]]`
- ⑦ `html <- read_html(htmlFull)`
- ⑧ `newsList <- html %>% html_nodes("#section_body ul")`

Name	Type	Value
newsList	list [4] (S3: xml_nodeset)	List of length 4
[[1]]	list [2] (S3: xml_node)	List of length 2
	list [2] (S3: xml_node)	List of length 2
	list [2] (S3: xml_node)	List of length 2
	list [2] (S3: xml_node)	List of length 2
	list [2] (S3: xml_node)	List of length 2
	list [2] (S3: xml_node)	List of length 2
(xml attributes)	character [1]	'type06_headline'
[[2]]	list [2] (S3: xml_node)	List of length 2
[[3]]	list [2] (S3: xml_node)	List of length 2
[[4]]	list [2] (S3: xml_node)	List of length 2

텍스트 데이터 존재

1. 웹 크롤링 실습2

RSelenium

실습3. RSelenium을 활용한 크롤링 만들기

- ⑨ for(one in newsList) {
- ⑩ temp <- one %>% html_nodes("li dl dt")
 %>% html_nodes("a")
 %>% html_attr("href")
- ⑪ print(temp)
- ⑫ }

```
[1] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=092&aid=0002157054"  
[2] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=092&aid=0002157054"  
[3] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=448&aid=0000267714"  
[4] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=448&aid=0000267714"  
[5] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=018&aid=0004318933"  
[6] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=018&aid=0004318933"  
[7] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=018&aid=0004318903"  
[8] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=018&aid=0004318903"  
[9] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=008&aid=0004180234"  
[10] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=008&aid=0004180234"  
[1] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=020&aid=0003201030"  
[2] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=020&aid=0003201030"  
[3] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=366&aid=0000427833"  
[4] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=366&aid=0000427833"  
[5] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=028&aid=0002444474"  
[6] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=028&aid=0002444474"  
[7] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=015&aid=0004099319"  
[8] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=015&aid=0004099319"  
[9] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=029&aid=0002510954"  
[10] "/main/read.nhn?mode=LSD&mid=shm&sid1=105&oid=029&aid=0002510954"
```

Q & A