

Lecture Note 06

Data Collection I



Dohyung Bang

Fall, 2021

Syllabus

Week	Date	Topic	Note
1	9/6(월)	R Basic - R 기초 문법 학습	
2	9/13(월)	R Basic – Data Manipulation I	과제#1
3	9/20(월) (추석)	<추석> (보충영상) R Basic - Data Manipulation II	
4	9/27(월)	Descriptive Analytics I - 데이터 요약하기/상관관계/차이검증	과제#2
5	10/4(월) (대체공휴일)	<대체공휴일> (보충영상) Descriptive Analytics II - 데이터 시각화	과제#2
6	10/11(월) (대체공휴일)	<대체공휴일> (보충영상) Supplementary Topic I - 외부 데이터 수집 (정적 콘텐츠 수집)	과제#4 과제#3
7	10/18(월)	Predictive Analytics I – Linear regression & Logistic Regression	
8	10/25(월)	Predictive Analytics II – Clustering & Latent Class Analysis	시험 대체 수업
9	11/1(월)	Predictive Analytics III – Tree-based Model and Bagging (Random Forest)	과제#4
10	11/8(월)	Predictive Analytics IV – Association Rules	
11	11/15(월)	Supplementary Topic II - 외부 데이터 수집 (동적 콘텐츠 수집)	
12	11/22(월)	Prescriptive Analytics I – Linear Programming	과제#5
13	11/29(월)	Prescriptive Analytics II – Data Envelopment Analysis (DEA)	
14	12/6(월)	Prescriptive Analytics III – Integer Programming	과제#6
15	12/13(월)	Prescriptive Analytics IV – Simulation	Quiz
16	12/20(월)	Final Presentation	

※ 사전 준비사항 1) 크롬 설치

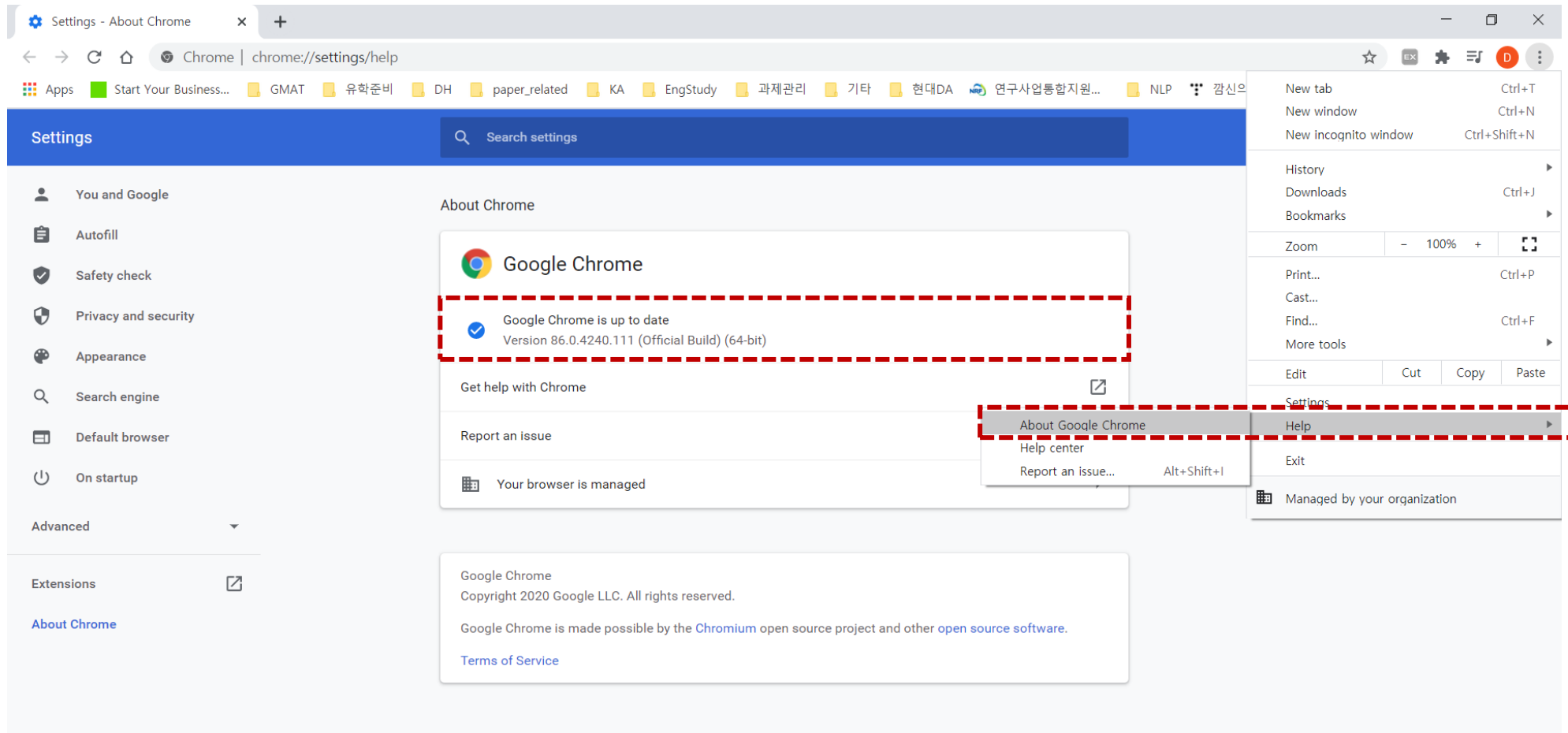


수집기는 기본적으로 크롬 환경으로 구성할 예정이므로
“크롬(Chrome)” 브라우저 최신버전 다운

<https://www.google.com/chrome/>

※ 사전 준비사항 2) 크롬 드라이버 설치

1) 크롬의 버전 정보를 확인



※ 사전 준비사항 2) 크롬 드라이버 설치

2) 아래 링크에서 크롬과 동일한 버전의 크롬 드라이버를 설치

<https://sites.google.com/a/chromium.org/chromedriver/>

ChromeDriver - WebDriver for Chrome

CHROMEDRIVER

CAPABILITIES & CHROMEOPTIONS

CHROME EXTENSIONS

CHROMEDRIVER CANARY

CONTRIBUTING

▼ DOWNLOADS

VERSION SELECTION

▼ GETTING STARTED

ANDROID

CHROMEOS

▼ LOGGING

PERFORMANCE LOG

MOBILE EMULATION

▼ NEED HELP?

CHROME DOESN'T START OR CRASHES IMMEDIATELY

CHROMEDRIVER CRASHES

CLICKING ISSUES

KEYBOARD SUPPORT

ChromeDriver

WebDriver is an open source tool for automated testing of webapps across many browsers. It provides capabilities for navigating to web pages, user input, JavaScript execution, and more. ChromeDriver is a standalone server that implements the [W3C WebDriver standard](#). ChromeDriver is available for Chrome on Android and Chrome on Desktop (Mac, Linux, Windows and ChromeOS).

You can view the current implementation status of the WebDriver standard [here](#).

All versions available in [Downloads](#)

- Latest stable release: [ChromeDriver 86.0.4240.22](#)
- Latest beta release: [ChromeDriver 87.0.4280.20](#)





ChromeDriver Documentation


- [Getting started with ChromeDriver on Desktop](#) (Windows, Mac, Linux)
 - [ChromeDriver with Android](#)
 - [ChromeDriver with ChromeOS](#)
- [ChromeOptions](#), the capabilities of ChromeDriver
- [Mobile emulation](#)

※ 사전 준비사항 3) Java 설치

1) 아래 링크에서 운영체제에 맞는 Java 설치(Win 64비트는 반드시 오프라인(64비트)로 설치해야 함)

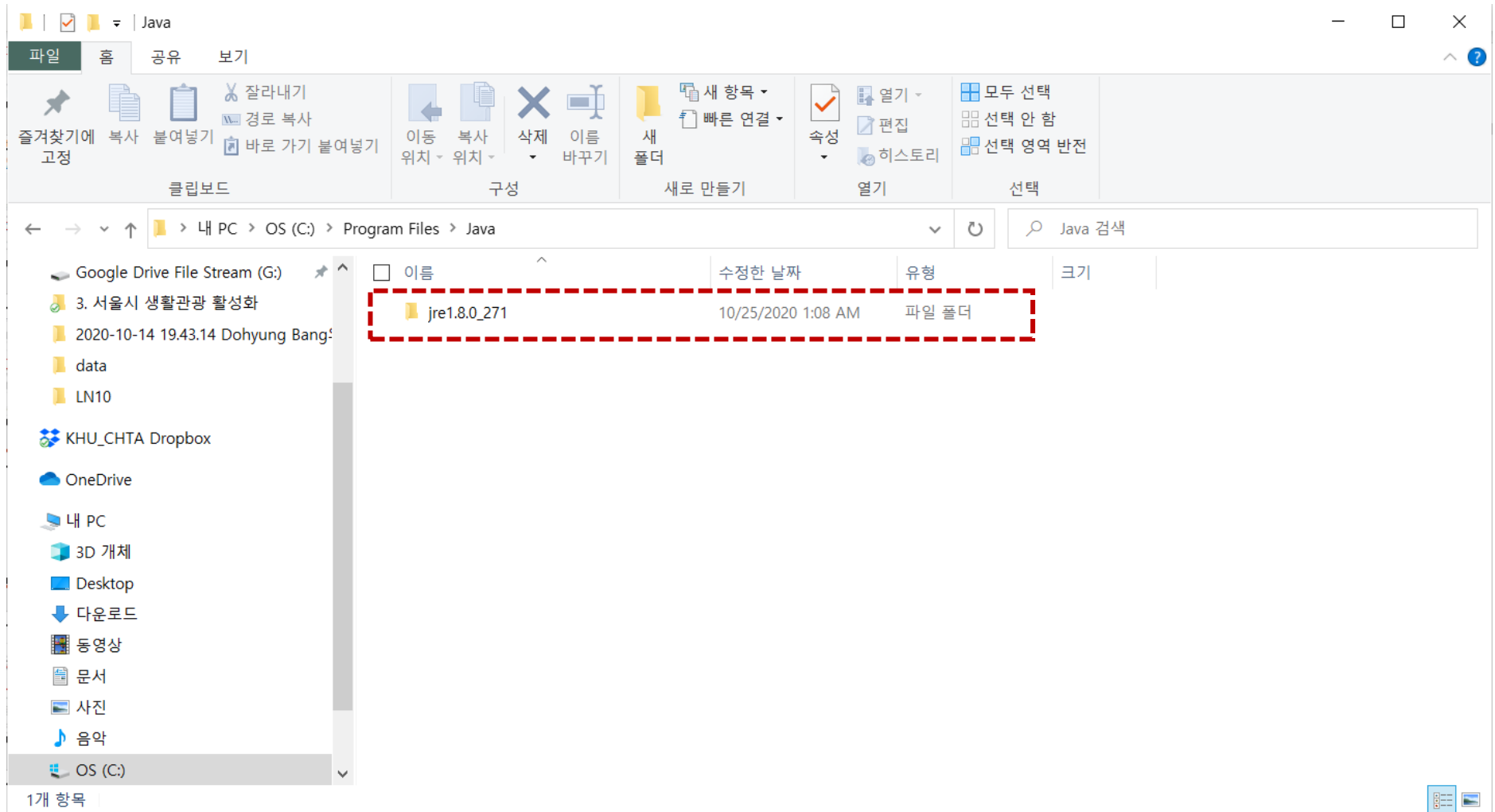
<https://www.java.com/ko/download/manual.jsp>

Windows  무엇을 선택해야 할까요?			
	Windows 온라인 파일 크기: 1.98 MB	지침	Java를 설치한 후 브라우저에서 Java를 사용으로 설정하려면 브라우저를 재시작해야 할 수 있습니다.
	Windows 오프라인 파일 크기: 69.53 MB	지침	
	Windows 오프라인 (64비트) 파일 크기: 79.5 MB	지침	
32비트 및 64비트 브라우저를 교대로 사용하는 경우, 각 브라우저에 대해 Java Plug-in이 필요하므로 32비트 Java와 64비트 Java를 모두 설치해야 합니다. » Windows용 64비트 Java에 대한 FAQ			

Mac OS X Mac FAQ			
	Mac OS X (10.7.3 버전 이상) 파일 크기: 80.75 MB	지침	Java를 설치한 후 브라우저에서 Java를 사용으로 설정하려면 브라우저를 재시작해야 할 수 있습니다.
* Oracle Java(버전 7 및 이후 버전)를 설치하고 실행하려면 Mac OS X 10.7.3(Lion) 이상을 실행하는 Intel 기반 Mac과 설치용 관리자 권한이 필요합니다. » 자세한 내용			

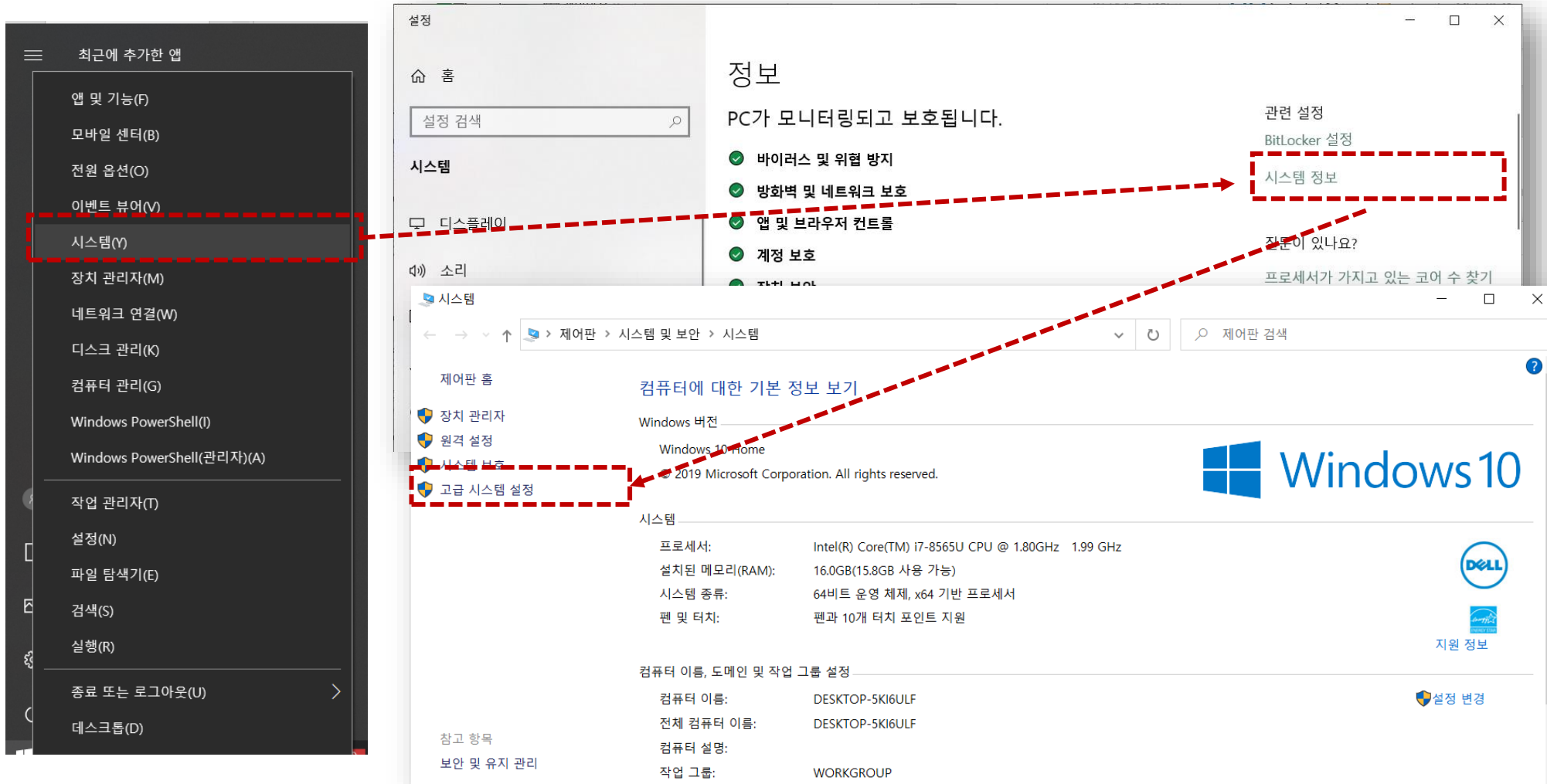
※ 사전 준비사항 3) Java 설치

2) 설치된 경로 확인



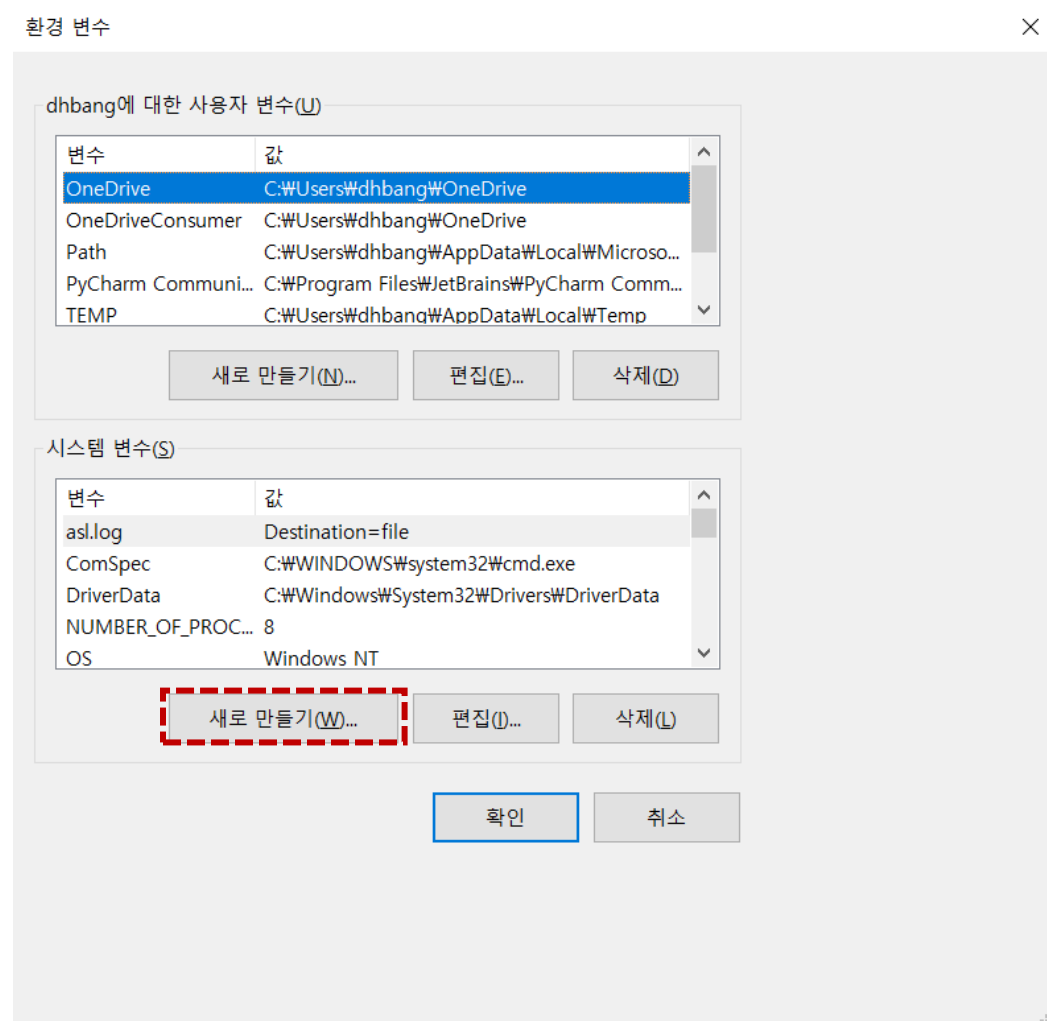
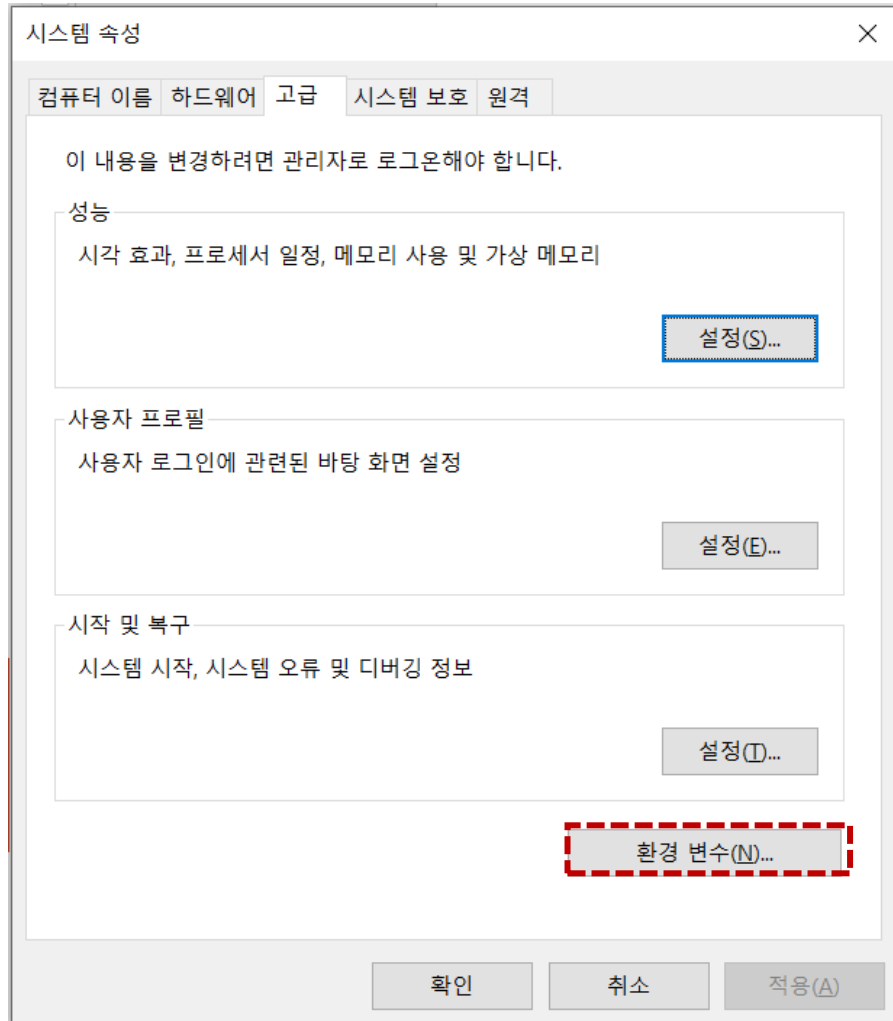
※ 사전 준비사항 3) Java 설치

3) Java는 설치 후 환경변수로 설정해 주어야 함



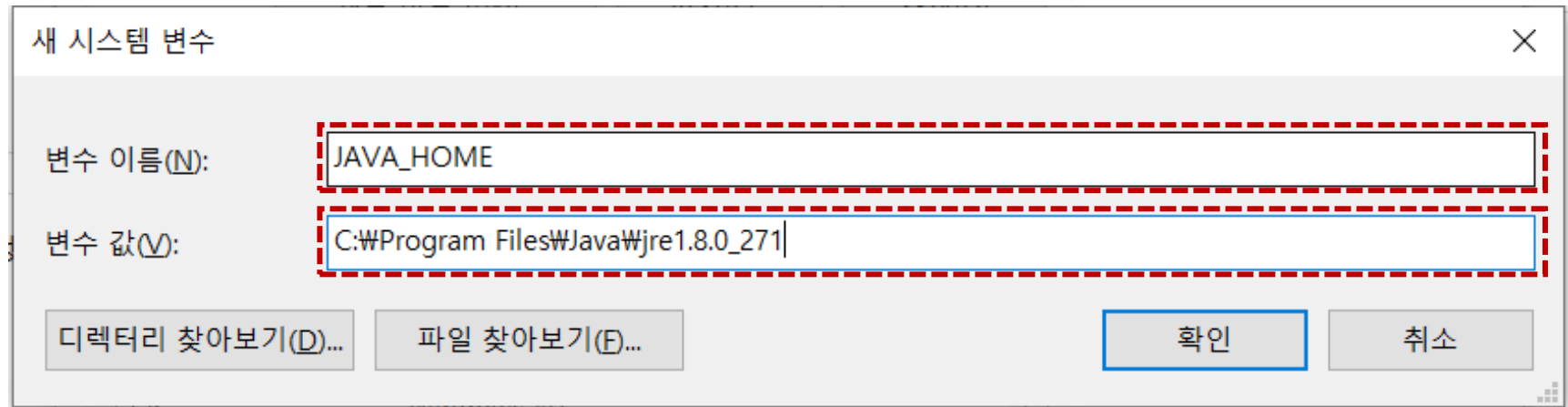
※ 사전 준비사항 3) Java 설치

3) Java는 설치 후 환경변수로 설정해 주어야 함 – 시스템 변수 수정



※ 사전 준비사항 3) Java 설치

3) Java는 설치 후 환경변수로 설정해 주어야 함 – 시스템 변수 수정



※ 사전 준비사항 3) Java 설치

3) Java는 설치 후 환경변수로 설정해 주어야 함 – 사용자 변수 수정

환경 변수

dhbang에 대한 사용자 변수(U)

변수	값
OneDrive	C:\Users\dhbang\OneDrive
OneDrive Consumer...	C:\Users\dhbang\OneDrive
Path	C:\Users\dhbang\AppData\Local\Microso...
PyCharm Communi...	C:\Program Files\JetBrains\PyCharm Comm...
TEMP	C:\Users\dhbang\AppData\Local\Temp

새로 만들기(N)... 편집(E)... 삭제(D)

시스템 변수(S)

변수	값
asl.log	Destination=file
ComSpec	C:\WINDOWS\system32\cmd.exe
DriverData	C:\Windows\System32\Drivers\DriverData
JAVA_HOME	C:\Program Files\Java\jre1.8.0_271
NUMBER OF PROC...	8

새로 만들기(W)... 편집(I)... 삭제(L)

확인 취소

환경 변수 편집

세미콜론(:) 뒤에 JAVA_HOME 경로를 추가

ta\Local\Microsoft\WindowsApps; C:\Program Files\Java\jre1.8.0_271
%PyCharm Community Edition%

새로 만들기(N) 편집(E) 찾아보기(B)... 삭제(D) 위로 이동(U) 아래로 이동(O) 텍스트 편집(T)...

확인 취소






























※ 사전 준비사항 3) Java 설치

4) 최종적으로 R 환경 상에서 Java 경로를 지정해주어야 끝이 남














```
12 ▾ # Java 경로 지정
13
14 ▾ ```{r}
15 Sys.setenv(JAVA_HOME = "C:/Program Files/Java/jre1.8.0_271")
16 ▴ ```
```

※ 사전 준비사항 4) Selenium 설치

<http://selenium-release.storage.googleapis.com/index.html>

	3.0-beta4	-	-
	3.0	-	-
	3.1	-	-
	3.10	-	-
	3.11	-	-
	3.12	-	-
	3.13	-	-
	3.14	-	-
	3.14.1	-	-
	3.150	-	-
	3.2	-	-
	3.3	-	-
	3.4	-	-
	3.5	-	-
	3.6	-	-
	3.7	-	-
	3.8	-	-
	3.9.0	-	-
	3.9	-	-
	4.0-alpha-5	-	-
	4.0-alpha-6	-	-
	4.0-alpha	-	-
	4.0-alpha1	-	-
	4.0-alpha2	-	-
	4.0-alpha4	-	-
	4.0-alpha5	-	-
	4.0	-	-
	icons	-	-
	index.html	2014-01-13 22:12:39	0.01MB 704b0f841aad1b1428481b7ff3c759c0

Index of /4.0/

	Name	Last modified	Size	ETag
	Parent Directory	-	-	-
	selenium-html-runner-4.0.0-alpha-1.jar	2019-04-24 16:17:02	13.52MB	2eca35318710f46d1ba5ed5543a906c9
	selenium-html-runner-4.0.0-alpha-2.jar	2019-07-01 21:32:41	13.76MB	346d72e4f425bfec91c7073a46c96208
	selenium-java-4.0.0-alpha-1.zip	2019-04-24 16:17:01	8.46MB	db9ed262a07c1cd2bb6098263c7f1e7b
	selenium-java-4.0.0-alpha-2.zip	2019-07-01 21:32:33	8.74MB	2d31929580c3d829197eea97ade5f4f0
	selenium-server-4.0.0-alpha-1.jar	2019-04-24 16:16:58	10.62MB	c32b1dd1c12cdb42b48f345d65d657fb
	selenium-server-4.0.0-alpha-1.zip	2019-04-24 16:16:59	10.20MB	7f0bc4bb4fc2a5a7f0a262f62bf782d3
	selenium-server-4.0.0-alpha-2.jar	2019-07-01 21:32:04	10.79MB	d0676e6b3ee508b48416aba603662573
	selenium-server-4.0.0-alpha-2.zip	2019-07-01 21:32:11	10.47MB	fb19d62db44a7b163f1fbc2ff9df9ffa
	selenium-server-standalone-4.0.0-alpha-1.jar	2019-04-24 16:17:00	11.98MB	ac553ec987d16d2af8c8e3ef9061772c
	selenium-server-standalone-4.0.0-alpha-1.zip	2019-04-24 16:17:00	11.77MB	1974b11f970bad6e15c84e3840ec3897
	selenium-server-standalone-4.0.0-alpha-2.jar	2019-07-01 21:32:19	12.33MB	d000d97d24389fde5bfb94f450ede780
	selenium-server-standalone-4.0.0-alpha-2.zip	2019-07-01 21:32:27	12.26MB	2466773c71eeddea02004371a5e32324

※ 사전 준비사항 5) geckodriver 설치

<https://github.com/mozilla/geckodriver/releases/tag/v0.17.0>

Fixed

- The SetWindowRect command now returns the WindowRect when it is done
- Use ASCII versions of array symbols to properly display them in the Windows command prompt
- Use `SessionNotCreated` error instead of `UnknownError` if there is no current session

Assets 8

 geckodriver-v0.17.0-arm7hf.tar.gz	2.13 MB
 geckodriver-v0.17.0-linux32.tar.gz	2.17 MB
 geckodriver-v0.17.0-linux64.tar.gz	2.16 MB
 geckodriver-v0.17.0-macos.tar.gz	1.32 MB
 geckodriver-v0.17.0-win32.zip	2.15 MB
 geckodriver-v0.17.0-win64.zip	2.09 MB
 Source code (zip)	
 Source code (tar.gz)	

※ 사전 준비사항 6) c드라이브에 r_selenium 폴더 생성

The screenshot shows a Windows File Explorer window titled 'r_selenium'. The address bar indicates the path '내 PC > OS (C:) > r_selenium'. The left sidebar shows the 'OS (C:)' drive selected. The main pane displays a list of files and folders. A red dashed box highlights the search results for 'r_selenium 검색'.

이름	수정한 날짜	유형	크기
chromedriver	10/25/2020 1:36 AM	응용 프로그램	9,494KB
chromedriver_win32	10/25/2020 1:36 AM	압축(ZIP) 폴더	5,117KB
geckodriver	10/25/2020 1:26 AM	응용 프로그램	5,999KB
geckodriver-v0.17.0-win64	10/25/2020 12:46 AM	압축(ZIP) 폴더	2,137KB
selenium-server-standalone-4.0.0-alpha-1	10/25/2020 1:25 AM	Executable Jar File	12,271KB

Lecture 1-1

데이터 수집
접근 방법

웹 크롤링(Crawling) 혹은 스크래핑(Scrapping)이란?

정형화된 데이터든, 비정형화된 데이터든 웹 페이지에 노출된 정보를 필요한 형태로 선택, 추출 및 수집하는 행위를 통칭함



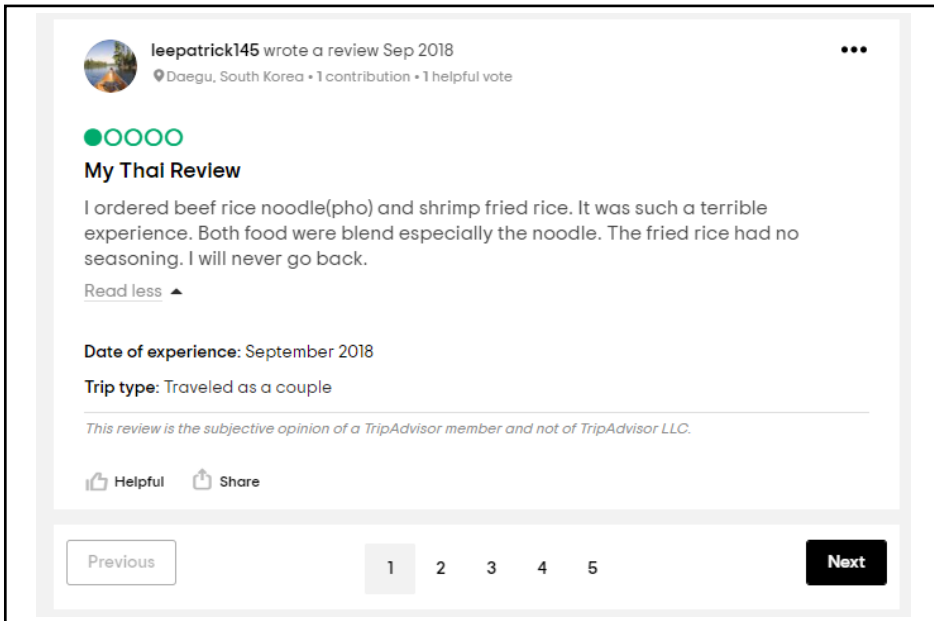
크롤링
(Crawling)

스크래핑
(Scrapping)

웹 페이지에서 정보를 추출해야 하므로
데이터 수집을 위해서는 웹 페이지에 대한 약간의
구조 이해가 수반되어야 함

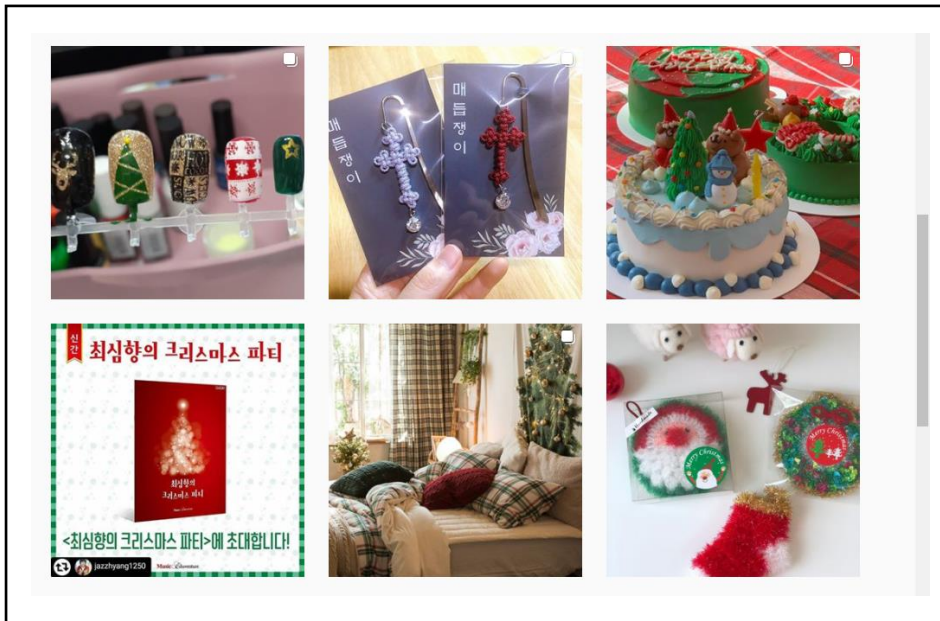
정적 콘텐츠 VS 동적 콘텐츠

정적 콘텐츠 (Static contents)



httr

동적/반응형 콘텐츠 (dynamic contents)



RSelenium

rvest

HTML(Hyper Text Markup Language)란 ?

- HTML (Contents & Structure) : 우리가 원하는 데이터를 포함하고 있는 핵심요소

```
<html>
<head>
  <title> HTML 예제 입니다. </title>
</head>
<body>
  <h1> 제목입니다. </h1>
  <p name = "my_name">
    에이치몰로 가고 싶으시면
    <a class = "my_link" href = "https://www.hyundaihmall.com/">링크</a>를
    클릭하시면 됩니다.
  </p>
</body>
</html>
```

- CSS (Presentation) : Font, Color, Border 등
- JavaScript (Behavior) : Display, Widget, Interactive Contents, Click 등

HTML(Hyper Text Markup Language)란 ?

- 꺾쇠(<>) 부분을 태그(Tag)라 부르며, “/”가 없으면 시작 태그, “/”가 있으면 종료태그임

```
<html>
<head>
  <title> HTML 예제 입니다. </title>
</head>
<body>
  <h1> 제목입니다. </h1>
  <p name = "my_name">
    에이치몰로 가고 싶으시면
    <a class = "my_link" href = "https://www.hyundaihmall.com/">링크</a>를
    클릭하시면 됩니다.
  </p>
</body>
</html>
```

시작태그(tag)

종료태그(tag)

속성명(attribute name)

속성값(attribute value)

우리가 원하는 데이터 !

HTML Elements 보는 방법

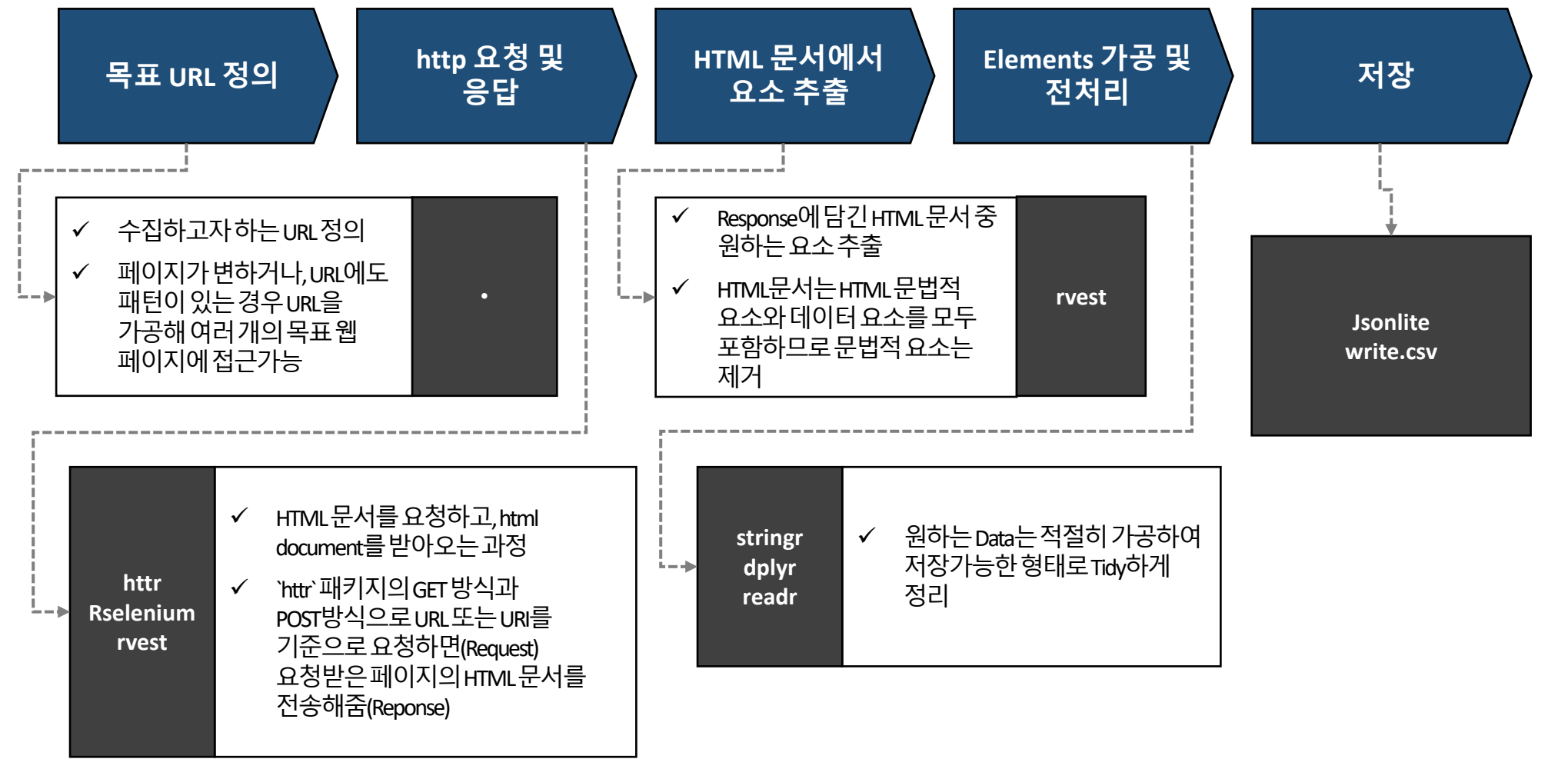
The image shows a screenshot of the Hmall website with a browser developer tool (Chrome DevTools) open, displaying the HTML structure of the page. The website features a purple header with the Hmall logo, a search bar, and navigation links. Below the header is a main banner area with a large image of two women and a smaller image of a man. The developer tool is open to the 'Elements' tab, showing the HTML structure of the page. The structure includes a header section with a search bar, a navigation bar with links like '카테고리', '로그인', and '회원가입', and a main content area with a large banner and a smaller banner. The HTML structure is as follows:

```

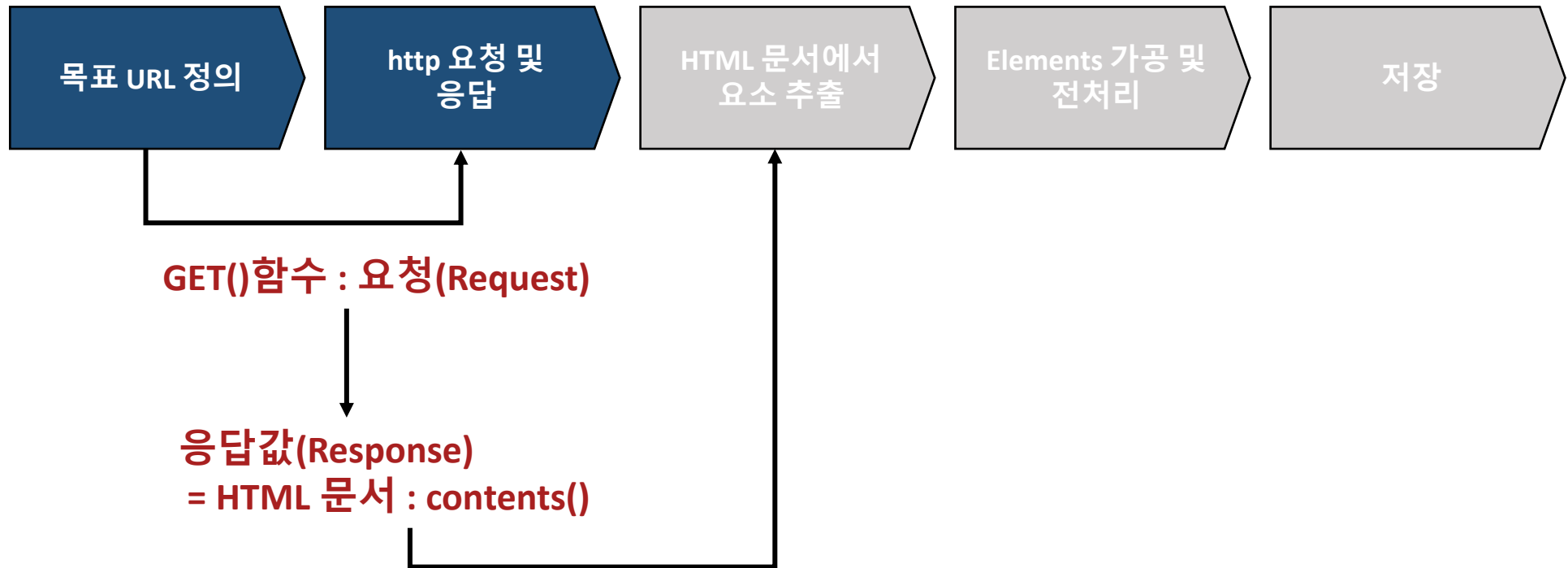
<!-- header -->
<div class="header">
  <div class="header-banner" style="background-color: rgb(132, 112, 255);">
    <span class="banner-content" id="top_bar"></span>
  </div>
  <div class="header-util">
    <span class="fl"></span>
    <span class="fr before-login"></span>
    <!-- 모바일 앱 다운로드 -->
    <form action="/CS/eva/evntTmplSmsSend.do" name="smsForm" method="post"></form>
  </div>
  <div class="header-title-area"></div>
  <div class="header-service-area"></div>
  <div class="content-quicklink-wrap">
    <ul class="quicklink ql-left id="main_tab">
      <li class="first"></li>
      <li></li>
      <li></li>
      <li></li>
      <li></li>
      <li></li>
      <li></li>
    </ul>
    <div class="visit-indicator"></div>
  </div>
  <form name="registMemberForm" method="post"></form>
</div>
<!-- header -->
<div class="content">
  <div id="content_wrap">
    <div class="middle-text-banner-bg" style="display: none; background-color: rgb(176, 163, 255);"></div>
    <div class="main-banner-container contrast-br" style="background-color: rgb(198, 185, 155);">
      <div class="main-banner-separator"></div>
      <div class="main-banner-wrap">
        <div class="main-banner">
          <div class="slider-container multiple">
            <div class="slider-overwrap">
              <ul class="main-banner-slide slider-initialized">
                <li data-name="HOT ISSUE" class="slide-child" data-index="0" data-child-index="4" style="width: 900px; position: relative; top: 0px; left: 0px; z-index: 99; opacity: 0; transition: opacity 500ms linear 0s;"></li>
                <li data-name="쇼핑에너지" class="slide-child" data-index="1" data-child-index="3" style="width: 900px; position: relative; top: 0px; left: -900px; z-index: 99; opacity: 0; transition: opacity 500ms linear 0s;"></li>
              </ul>
            </div>
          </div>
        </div>
      </div>
    </div>
  </div>
</div>

```

R을 이용한 웹 크롤링 절차



HTTP 요청 및 응답



URL과 URI의 차이

$$\underline{\text{URI}} = \text{URL} + \text{Query}$$

네이버 블로그 :

`https://section.blog.naver.com/BlogHome.nhn?directoryNo=0¤tPage=1&groupId=0`

URL

Query 시작 구분점

Query 분리 구분점

파라미터(Parameter)

네이버 부동산 :

`https://new.land.naver.com/complexes?ms=37.514592,127.105863,15&a=APT:ABYG:JGC&e=RETAIL`

Lecture 1-2

정적 콘텐츠
수집하기

rvest 패키지 이해하기

