

PYTHON TUTORING #7

School of Computing, KAIST & 대덕고등학교

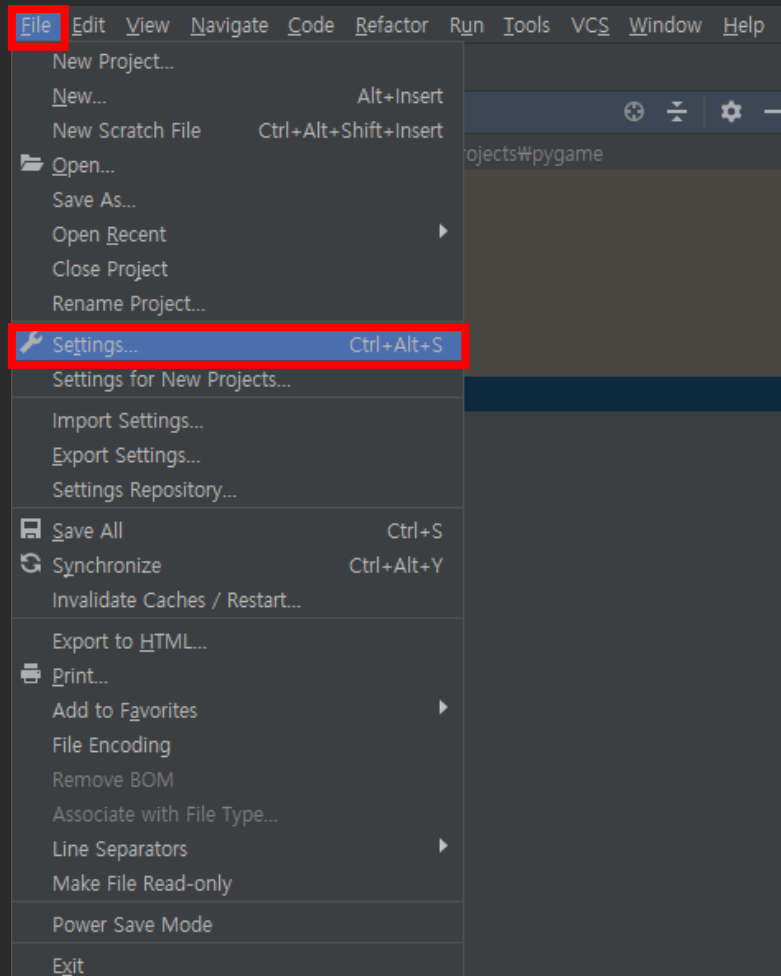
INTRO

- ① requests 라이브러리 설치
- ② 인터넷의 작동 방식
- ③ crawl.py
- ④ 서버의 Response와 HTML
- ⑤ 식단 이미지 수집하기

requests 라이브러리 설치

(1) 왼쪽 상단의
File 버튼 클릭 후,

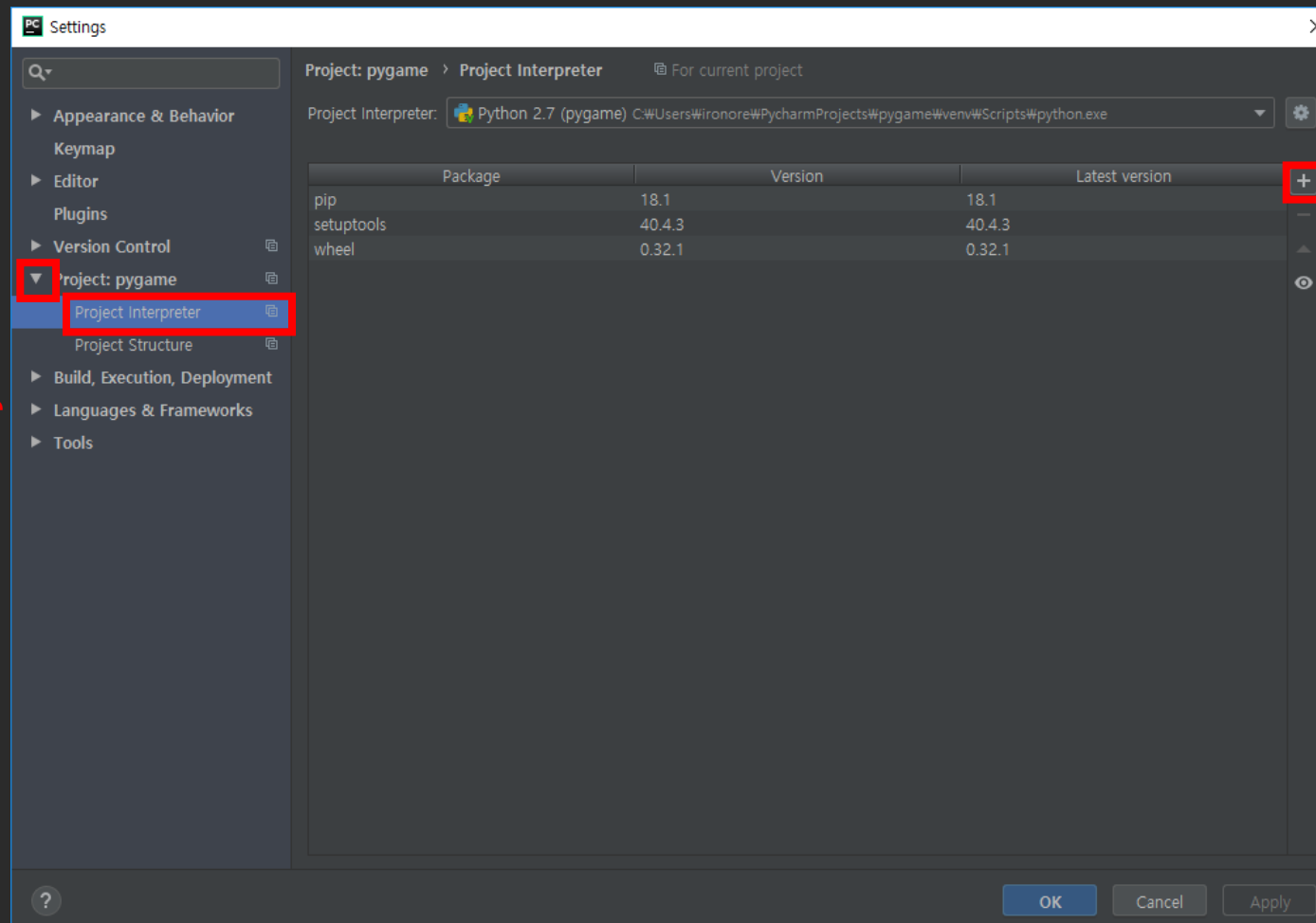
(2) Settings 버튼 클릭



requests 라이브러리 설치

(1) Project:name
왼쪽의 화살표 클릭

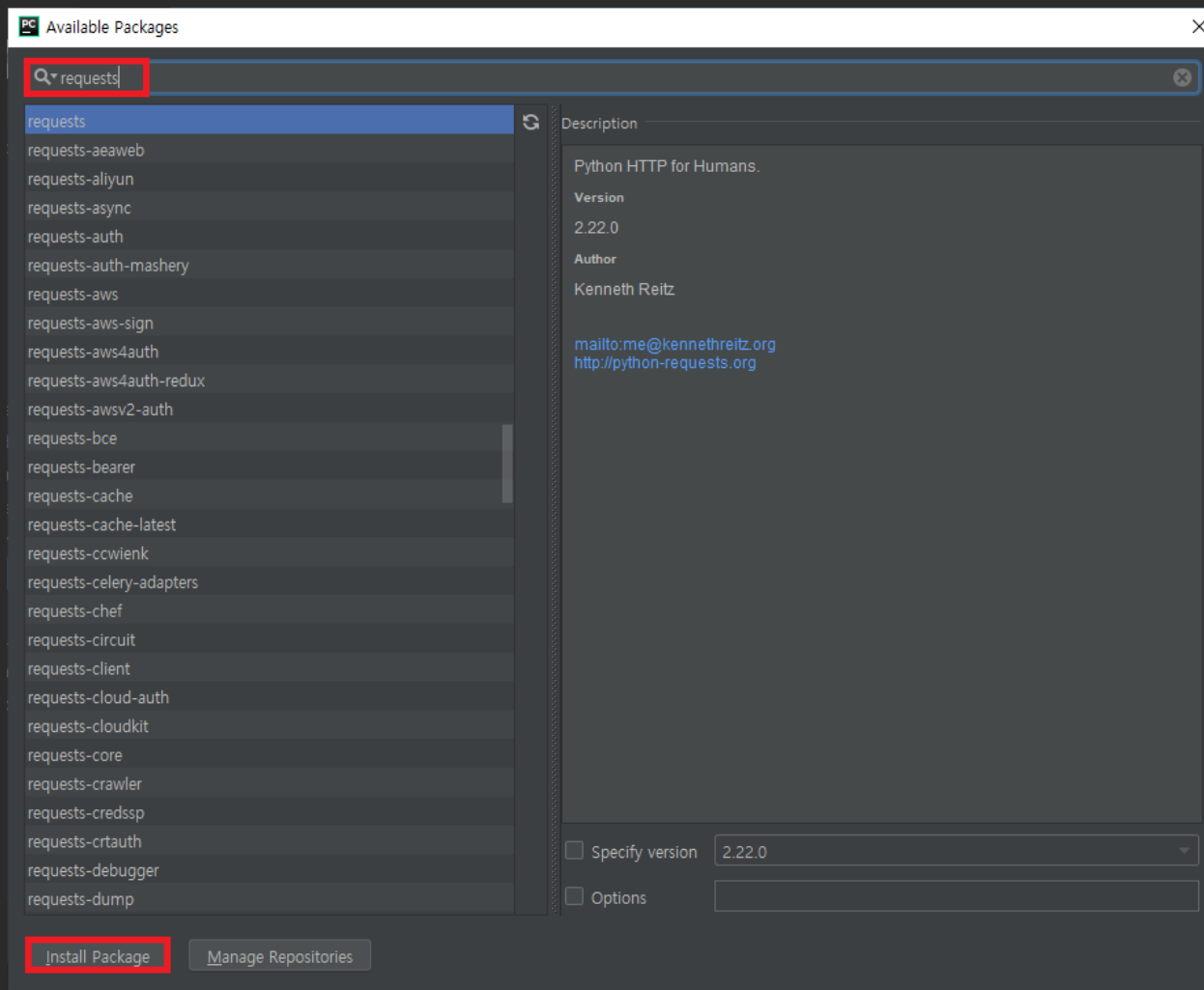
(2) Project Interpreter
버튼 클릭



(3) + 버튼 클릭

requests 라이브러리 설치

(1) 검색 칸에
requests 입력



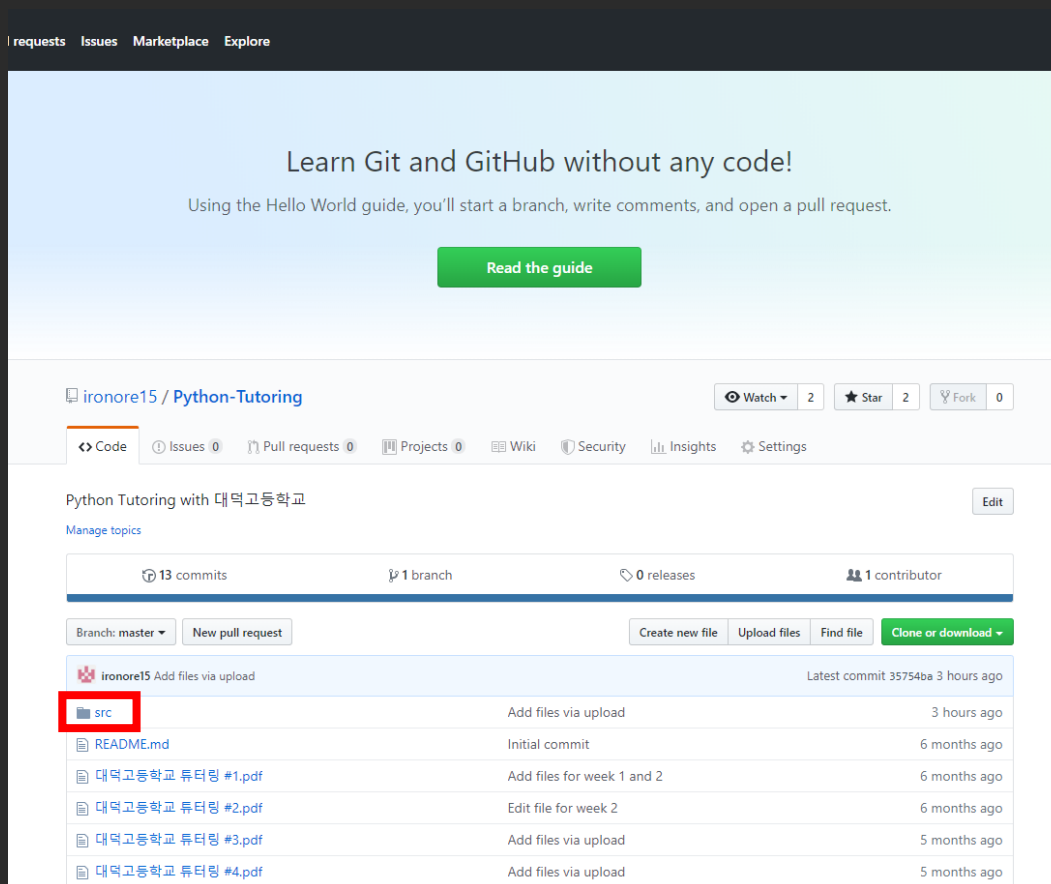
(2) Install Package
버튼 클릭 후 대기

여기서 잠깐!

- 라이브러리란?
- 미리 만들어 놓은 함수 또는 변수들을 사용하겠다는 것
- 프로그래밍 언어에서 자주 사용되지 않는 함수들은
- 라이브러리로 만들어 일부 필요한 사용자만 사용할 수 있게 만든 것
- 예시. `import math` 를 통해 `math`에 포함된 함수와 변수들을 사용할 수 있다.

프로그램 예제

<https://github.com/ironore15/Python-Tutoring>



requests Issues Marketplace Explore

Learn Git and GitHub without any code!
Using the Hello World guide, you'll start a branch, write comments, and open a pull request.

[Read the guide](#)

ironore15 / Python-Tutoring Watch 2 Star 2 Fork 0

Code Issues 0 Pull requests 0 Projects 0 Wiki Security Insights Settings

Python Tutoring with 대덕고등학교 [Edit](#)

Manage topics

13 commits 1 branch 0 releases 1 contributor

Branch: master [New pull request](#) [Create new file](#) [Upload files](#) [Find file](#) [Clone or download](#)

ironore15 Add files via upload Latest commit 35754ba 3 hours ago

src	Add files via upload	3 hours ago
README.md	Initial commit	6 months ago
대덕고등학교 튜터링 #1.pdf	Add files for week 1 and 2	6 months ago
대덕고등학교 튜터링 #2.pdf	Edit file for week 2	6 months ago
대덕고등학교 튜터링 #3.pdf	Add files via upload	5 months ago
대덕고등학교 튜터링 #4.pdf	Add files via upload	5 months ago

(1) src 버튼을 클릭 후,

(2) crawl.py 버튼을 누르면 오늘 사용할

프로그램 예제

<https://github.com/ironore15/Python-Tutoring>

```
ironore15 Update crawl.py 7439742 now
1 contributor

37 lines (28 sloc) 1.26 KB
Raw Blame History

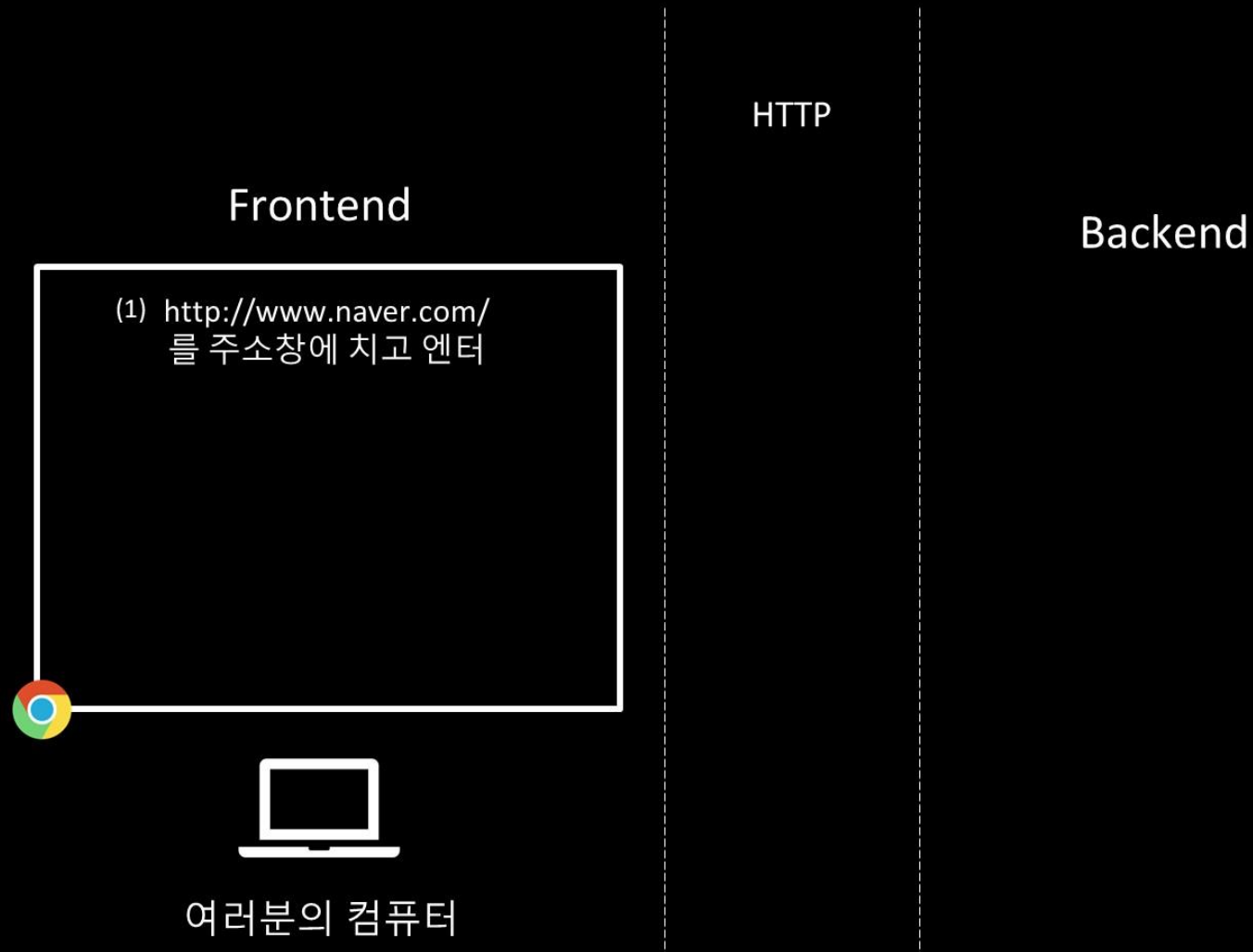
Code navigation is still being calculated for this commit. Check back in a bit. Beta Learn more or give us feedback

1 import requests
2
3 url = 'http://daedeokhs.djsch.kr'
4 path = '/schedule/list.do'
5 query = 's=taedokhs&schdYear=2019&schdMonth=11'
6 r = requests.get(url + path + '?' + query)
7
8 index = r.text.find('<th scope="row"')
9 while index != -1: # 검색에 실패할 때까지
10     final_index = r.text.find("</th>", index)
11     print(r.text[index:final_index+5])
12     next_index = r.text.find('<th scope="row"', index + 1) # find의 두 번째 인자는 start index
13     title_start = r.text.find('title=', index, next_index) # find의 세 번째 인자는 end index
14     if title_start != -1:
15         title_end = r.text.find('"', title_start + 7)
16         print(r.text[title_start+7:title_end])
17     index = next_index
18
19
20 path = '/boardCnts/list.do'
21 query = 'boardID=49529&m=040602&s=taedokhs'
22
23 r = requests.get(url + path + '?' + query)
24 print(r.text)
25
26 index = r.text.find('/upload/board/49529/')
27 while index != -1:
28     final_index = r.text.find('.jpg', index)
29     image_path = r.text[index:final_index+4]
30
31     image_path = image_path.replace('/thumb/', '/')
32     new_r = requests.get(url + image_path)
33     with open(str(index) + '.jpg', 'wb') as f:
34         f.write(new_r.content)
35
36     index = r.text.find('/upload/board/49529/', index + 1)
```

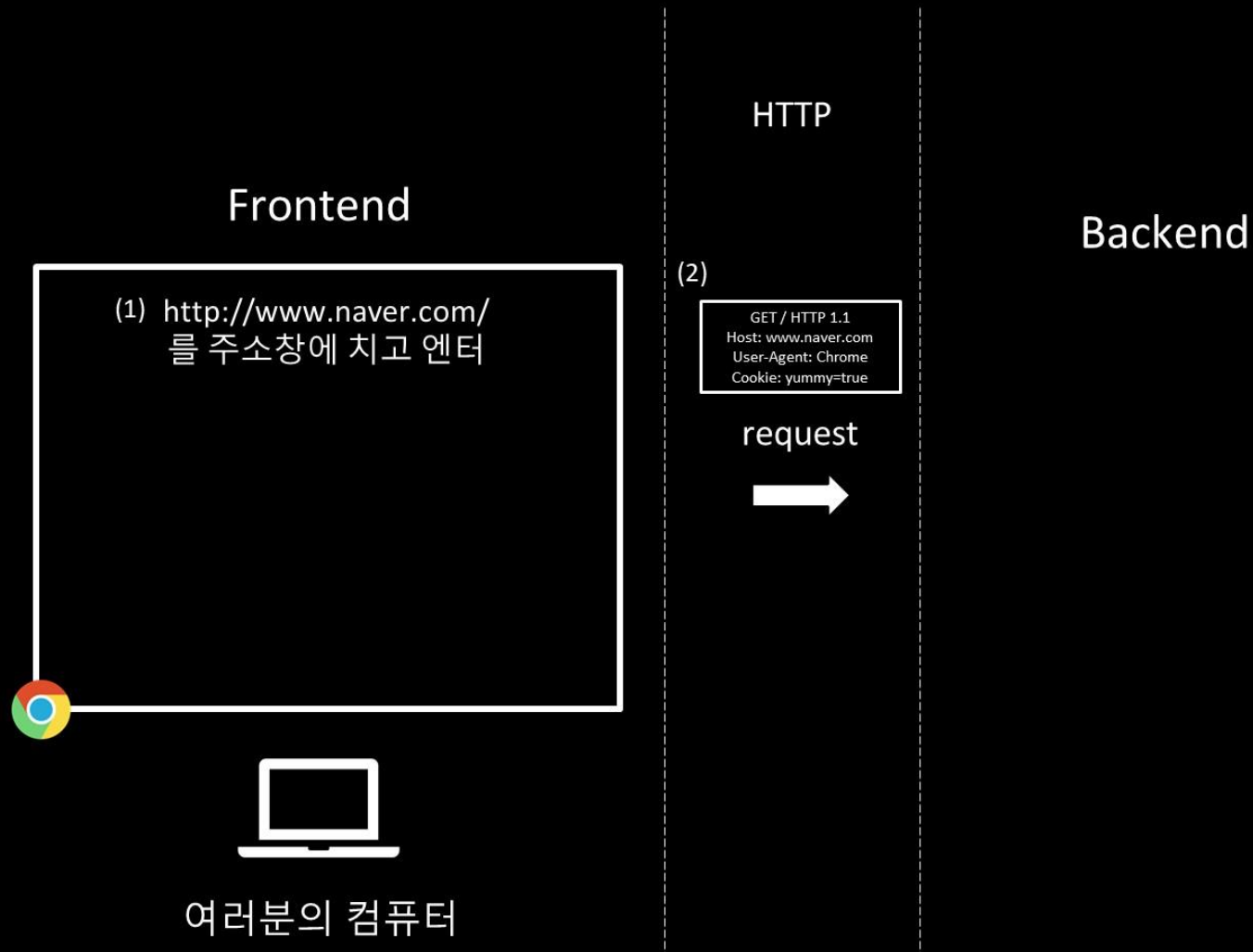
(1) src 버튼을 클릭 후,

(2) crawl.py 버튼을 누르면 오늘 사용할

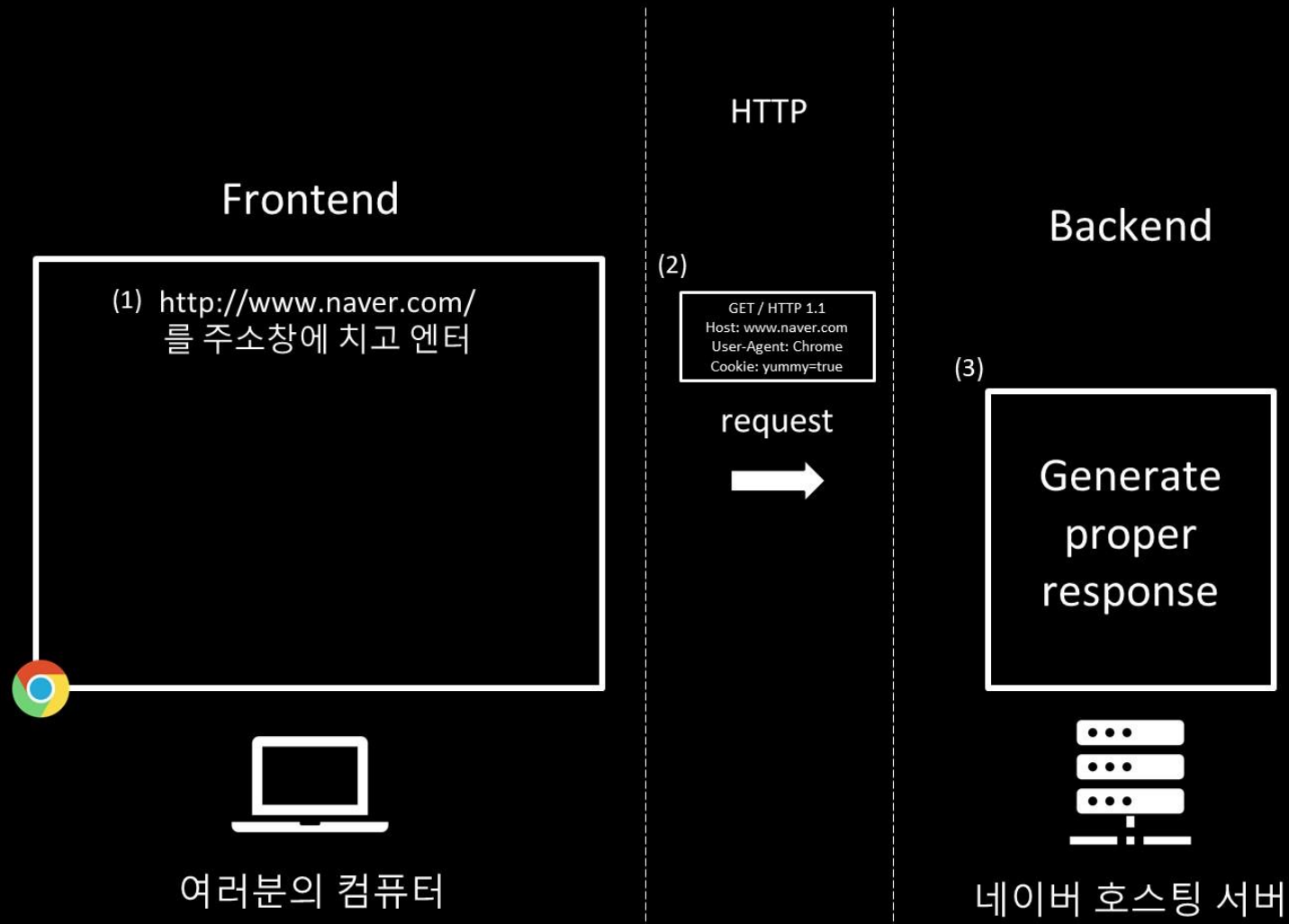
인터넷의 작동 방식

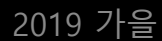


인터넷의 작동 방식

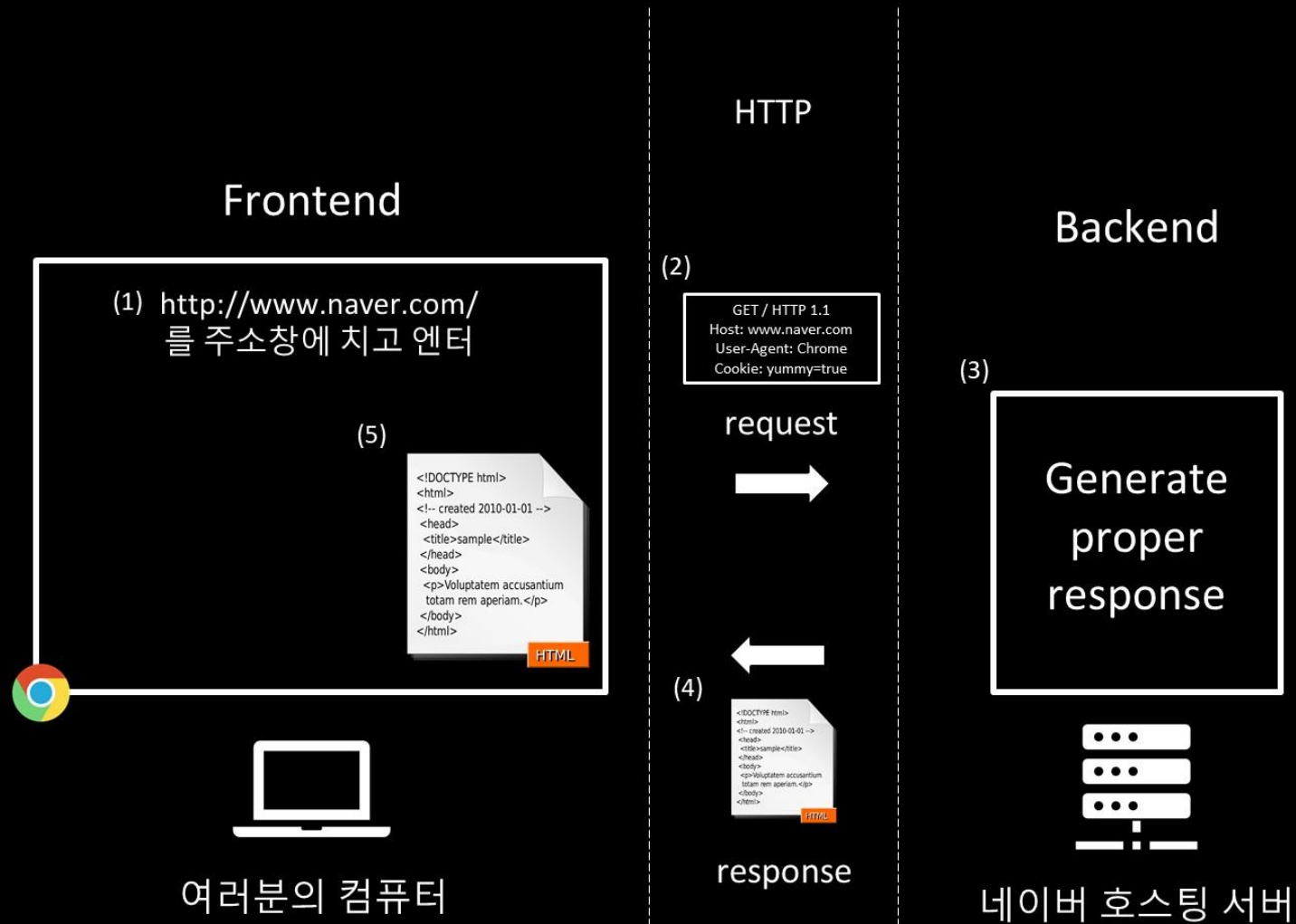


인터넷의 작동 방식

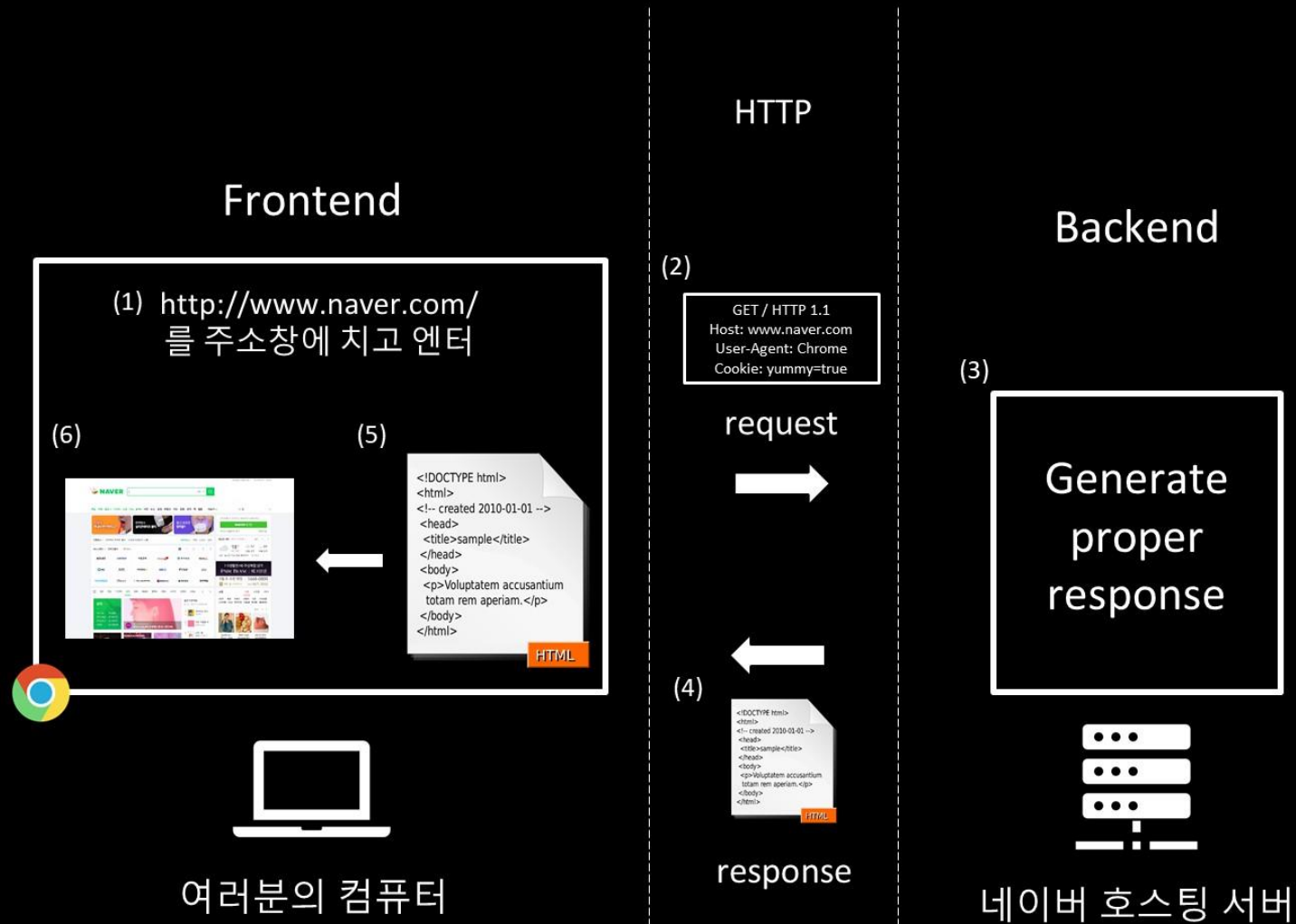




인터넷의 작동 방식

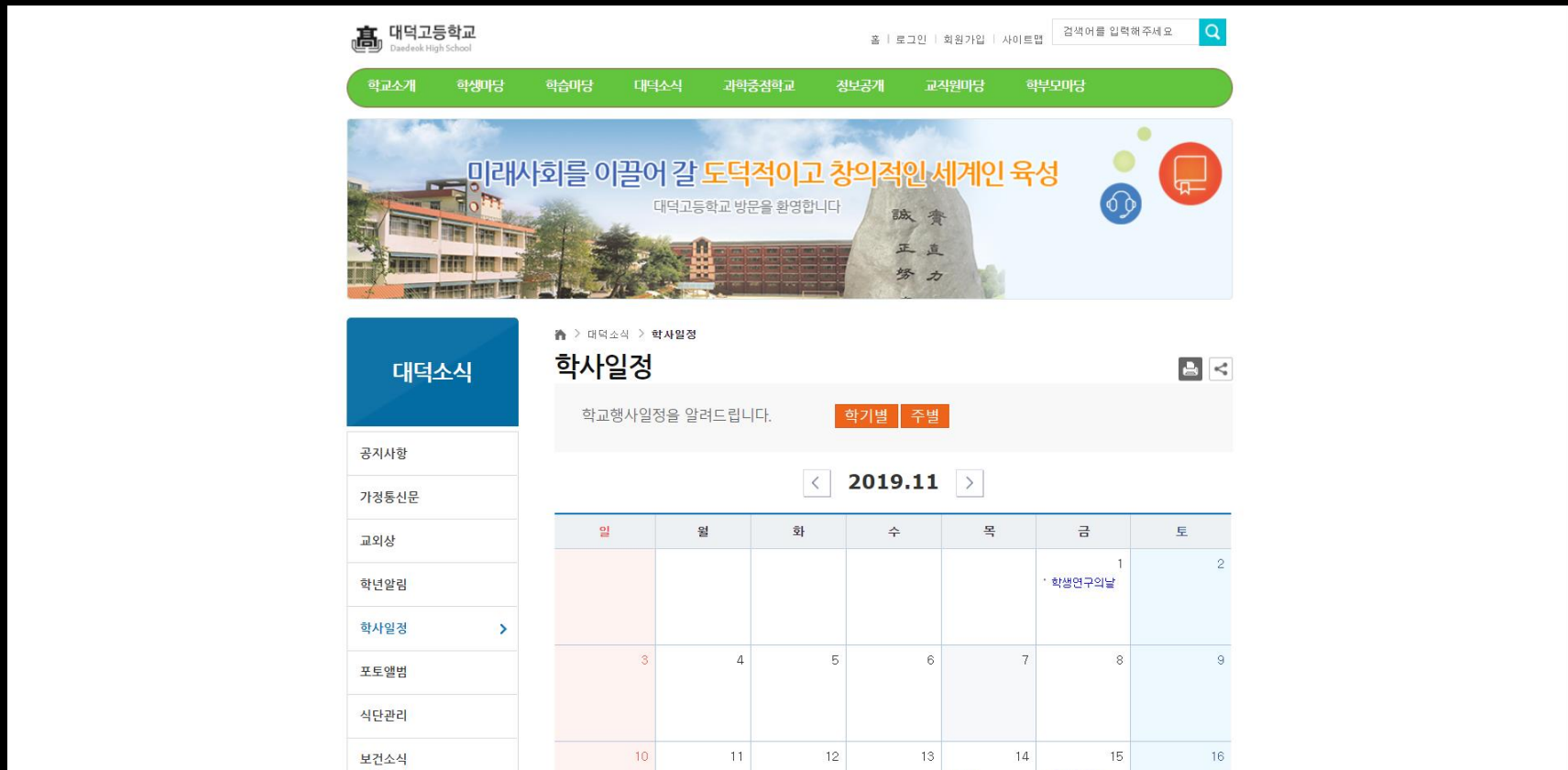


인터넷의 작동 방식



서버에게 보내는 정보

<http://daedeokhs.djsch.kr/schedule/list.do?s=taedokhs&schdYear=2019&schdMonth=11>



The screenshot shows the Daedeok High School website. The header includes the school logo and name, a search bar, and a navigation menu. The main banner features a school building and the text "미래사회를 이끌어갈 도덕적이고 창의적인 세계인 육성". Below the banner, there is a sidebar with links to various school information pages. The main content area displays the "학사일정" (Academic Calendar) for November 2019, showing a calendar grid with dates and specific events.

대덕고등학교
Daedeok High School

홈 | 로그인 | 회원가입 | 사이트맵 | 검색어를 입력해주세요

학교소개 | 학생마당 | 학습마당 | 대덕소식 | 과학중점학교 | 정보공개 | 교직원마당 | 학부모마당

미래사회를 이끌어갈 도덕적이고 창의적인 세계인 육성
대덕고등학교 방문을 환영합니다

대덕소식

공지사항
가정통신문
교외상
학년알림
학사일정
포토앨범
식단관리
보건소식

대덕소식 > 학사일정

학사일정

학교행사일정을 알려드립니다. 학기별 주별

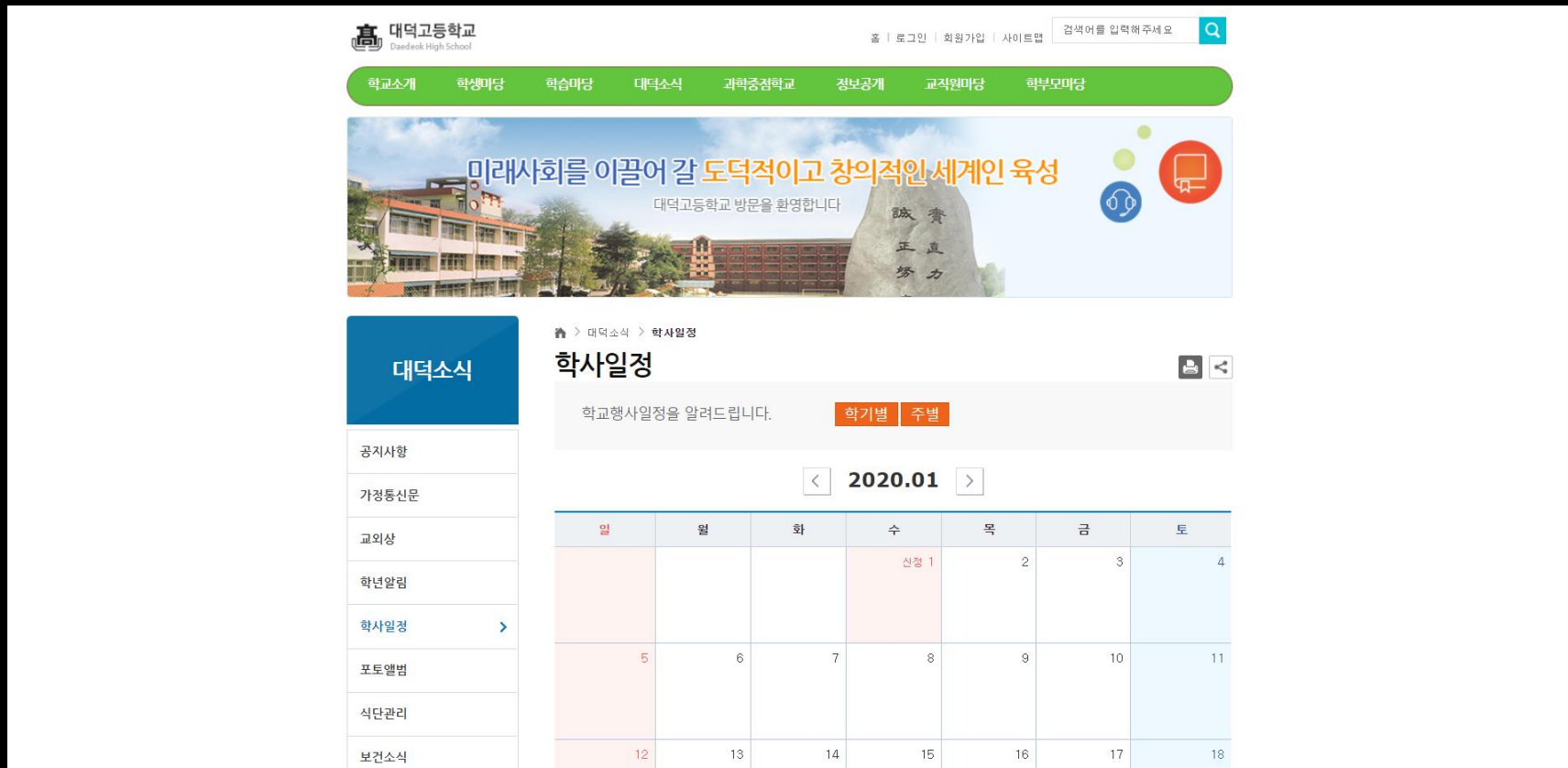
< 2019.11 >

일	월	화	수	목	금	토
					1 학생연구의날	2
3	4	5	6	7	8	9
10	11	12	13	14	15	16

서버에게 보내는 정보

<http://daedeokhs.djsch.kr/schedule/list.do?s=taedokhs&schdYear=2019&schdMonth=11>

<http://daedeokhs.djsch.kr/schedule/list.do?s=taedokhs&schdYear=2020&schdMonth=01>



High Daedeok High School

홈 | 로그인 | 회원가입 | 사이트맵 | 검색어를 입력해주세요

학교소개 학생매당 학습매당 대덕소식 과학중점학교 정보공개 교직원매당 학부모매당

미래사회를 이끌어갈 도덕적이고 창의적인 세계인 육성
대덕고등학교 방문을 환영합니다

대덕소식

공지사항
가정통신문
교외상
학년알림
학사일정
포토앨범
식단관리
보건소식

대덕소식 > 학사일정

학사일정

학교행사일정을 알려드립니다. 학기별 주별

< 2020.01 >

일	월	화	수	목	금	토
			신정 1	2	3	4
5	6	7	8	9	10	11
12	13	14	15	16	17	18

서버에게 보내는 정보

<http://daedeokhs.djsch.kr/schedule/list.do?s=taedokhs&schdYear=2019&schdMonth=12>

host

path

query

서버는 path와 query를 보고 알맞은 응답을 보낸다.

주석에 대해서...

```
1
2 # 샷 뒤에 적힌 내용은 주석으로, 코드에 영향을 주지 않음
3 # ppt에 적혀 있어도 똑같이 따라 적을 필요 X
4
5
6 """
7 따옴표 3개로 감싸진 줄들은 전부 주석 처리
8 test.py의 내용 전체를 실행시키기 싫다면,
9 실행을 원하지 않는 부분을 따옴표 3개로 감싸서
10 자신이 원하는 부분만 실행시키세요.
11 """
12
```

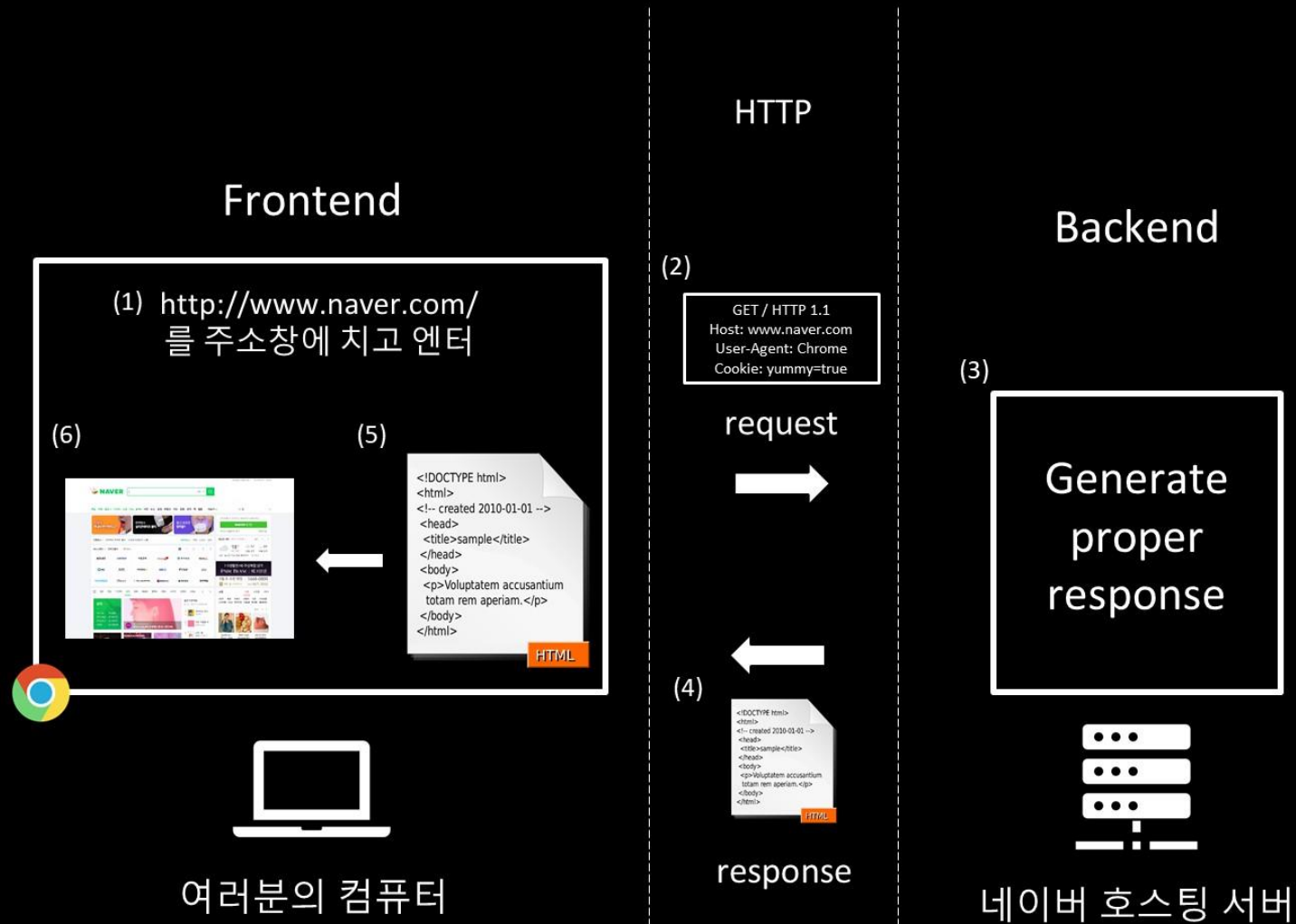
crawl.py

```
1 import requests
2
3
4 url = 'http://daedeokhs.djsch.kr'
5 path = '/schedule/list.do'
6 query = 's=taedokhs&schdYear=2019&schdMonth=11'
7
8 r = requests.get(url + path + '?' + query)
9
10
11 print(r.text)
12
```


crawl.py

```
1 .....
2 <!doctype html>
3 <html lang="ko">
4     <head>
5         <meta http-equiv="X-UA-Compatible" content="IE=edge">
6         <meta http-equiv="Content-Type" content="text/html; charset=utf-8" />
7         <meta name="viewport" content="width=device-width,initial-
8 scale=1.0,minimum-scale=1.0,maximum-scale=1.0" />
9         <title>대덕고등학교 홈페이지</title>
10        <!--[if lt IE 9]><script src="/js/html5.js"></script><![endif]-->
11 .....
12 .....
```

인터넷의 작동 방식



서버의 Response와 HTML




대덕고등학교
Daedeok High School

[홈](#) | [로그인](#) | [회원가입](#) | [사이트맵](#)



검색어를 입력해주세요

학교소개
학생마당
학습마당
대덕소식
과학중점학교
정보공개
교직원마당
학부모마당



미래사회를 이끌어 갈 도덕적이고 창의적인 세계인 육성

대덕고등학교 방문을 환영합니다

대덕소식

- 공지사항
- 가정통신문
- 교외상
- 학년알림
- 학사일정**
- 포토앨범
- 식단관리
- 보건소식

☞ > 대덕소식 > 학사일정

학사일정

학교행사일정을 알려드립니다.

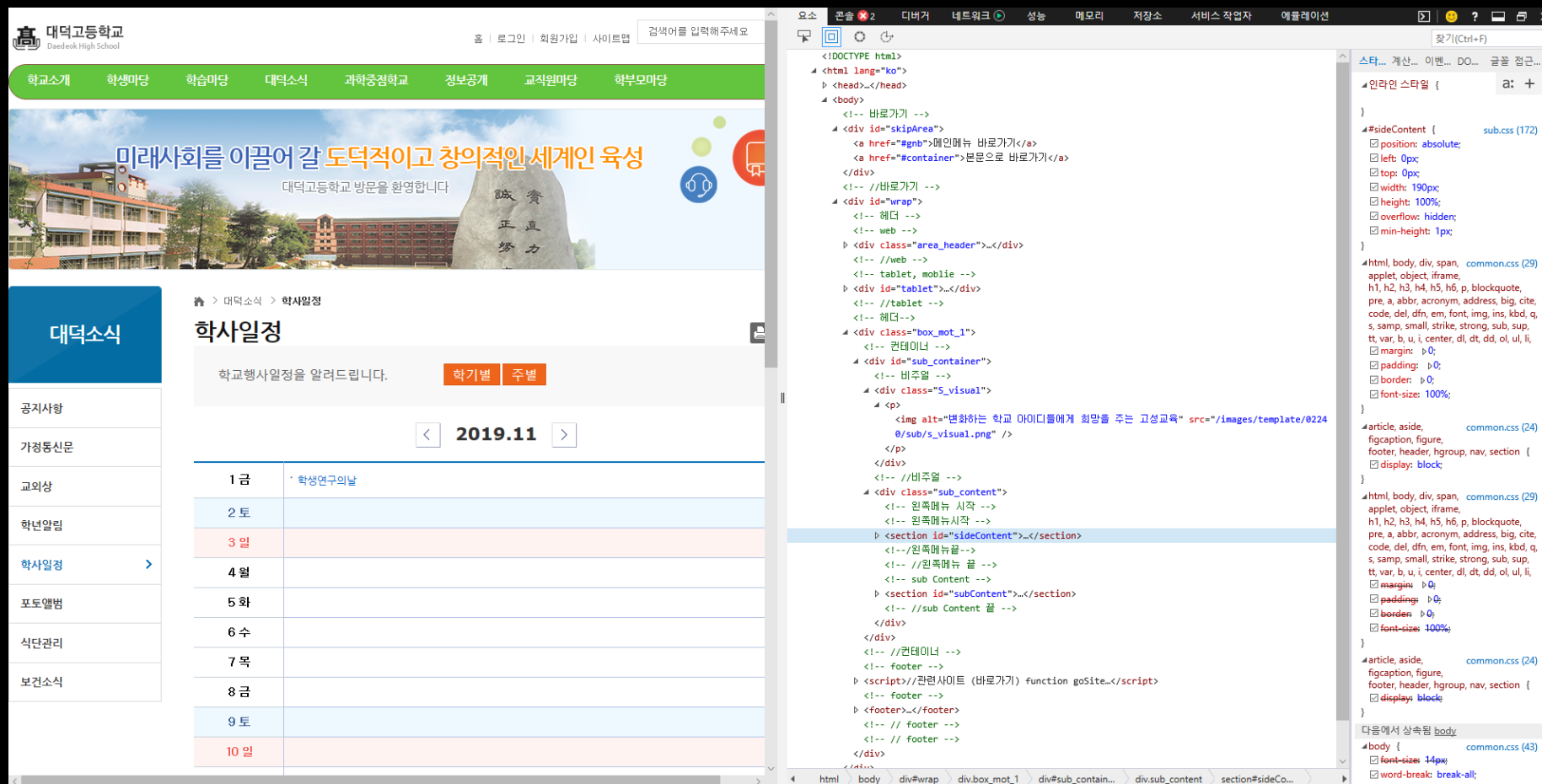
학기별

주별

<
2019.12
>

일	월	화	수	목	금	토
1	2	3	4	5	6	7
8	9	10 <small>* 기말고사(1,2)</small>	11 <small>* 기말고사(1,2)</small>	12 <small>* 기말고사(1,2)</small>	13 <small>* 기말고사(1,2)</small>	14
15	16	17	18	19	20	21
22	23	24 <small>* 학미교향악</small>	25 <small>* 성탄절</small>	26	27	28

서버의 Response와 HTML

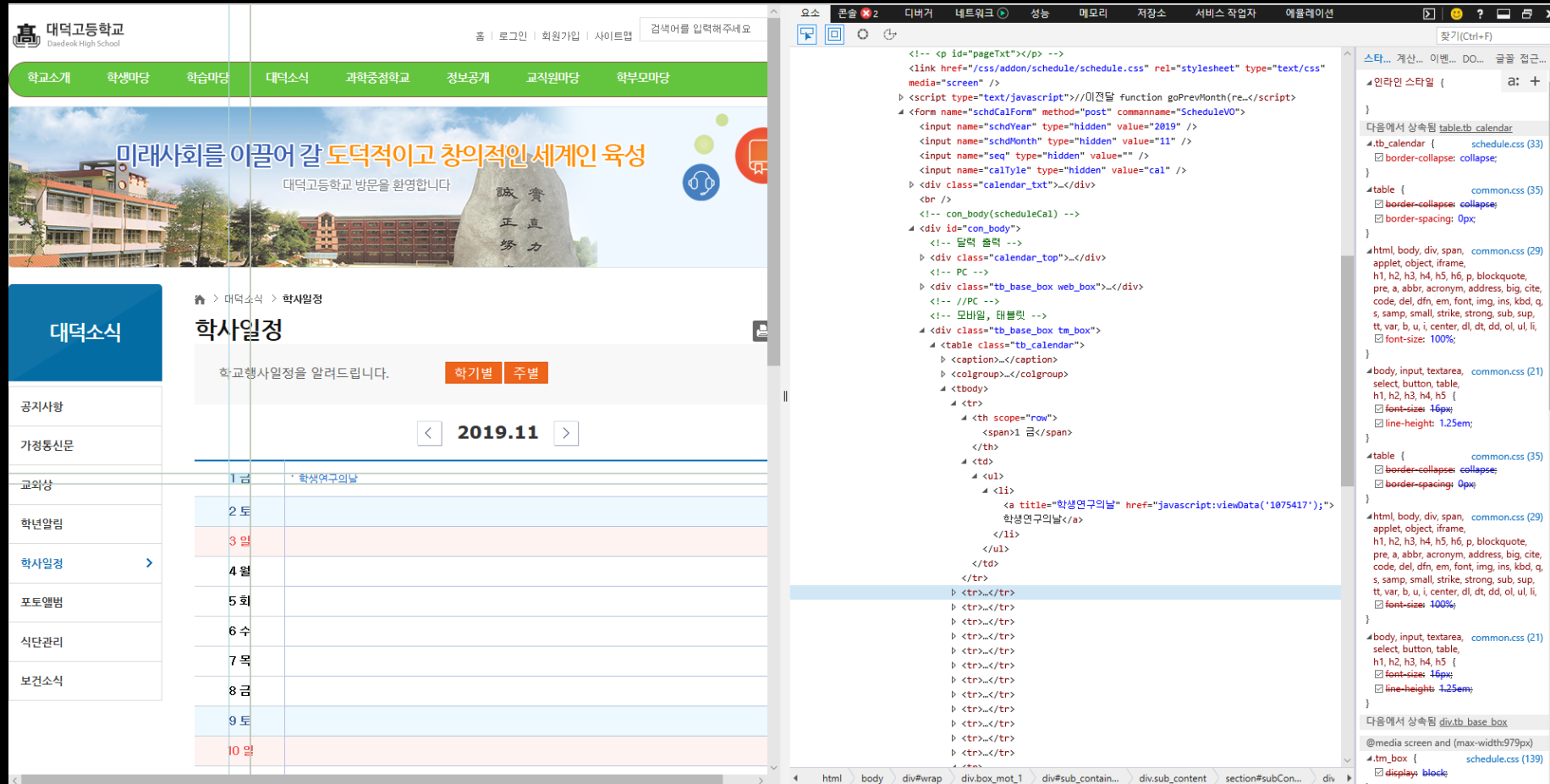


The screenshot shows a web browser displaying the Daedeok High School website. The developer tools are open, showing the HTML source code. The code includes a navigation bar, a main content area with a banner, and a sidebar with a menu. The HTML structure is as follows:

```
<!DOCTYPE html>
<html lang="ko">
<head>...</head>
<body>
  <!-- 바로가기 -->
  <div id="skipArea">
    <a href="#gnb">메인메뉴 바로가기</a>
    <a href="#container">본문으로 바로가기</a>
  </div>
  <!-- //바로그기 -->
  <div id="wrap">
    <!-- 헤더 -->
    <!-- web -->
    <div class="area_header">...</div>
    <!-- //web -->
    <!-- tablet, mobile -->
    <div id="tablet">...</div>
    <!-- //tablet -->
    <!-- 헤더 -->
    <div class="box_mot_1">
      <!-- 컨테이너 -->
      <div id="sub_container">
        <!-- 비주얼 -->
        <div class="S_visual">
          <p>
            
          </p>
        </div>
        <!-- //비주얼 -->
        <div class="sub_content">
          <!-- 왼쪽메뉴 시작 -->
          <!-- 왼쪽메뉴시작 -->
          <section id="sideContent">...</section>
          <!--//왼쪽메뉴끝-->
          <!-- //왼쪽메뉴 끝 -->
          <!-- sub Content -->
          <section id="subContent">...</section>
          <!-- //sub Content 끝 -->
        </div>
      </div>
      <!-- //컨테이너 -->
      <!-- footer -->
      <script>...</script>
      <!-- footer -->
      <footer>...</footer>
      <!-- // footer -->
      <!-- // footer -->
    </div>
  </div>
</body>
```

ctrl + shift + i 버튼으로
해당 웹 페이지의 HTML 코드를 볼 수 있다.

서버의 Response와 HTML



The screenshot shows a web browser displaying the 'Daedeok High School' (대덕고등학교) website. The page is titled '학사일정' (Academic Calendar) and shows a calendar for the year 2019. The calendar table is visible, with columns for the month and the day of the week. The source code is displayed in a developer tool, showing the HTML structure of the calendar table. The code includes a table with a caption '학사일정' and a table body with rows for each month. The table is styled with a blue header and a light blue body. The source code is shown in a developer tool, with the HTML structure of the calendar table highlighted. The code includes a table with a caption '학사일정' and a table body with rows for each month. The table is styled with a blue header and a light blue body. The source code is shown in a developer tool, with the HTML structure of the calendar table highlighted. The code includes a table with a caption '학사일정' and a table body with rows for each month. The table is styled with a blue header and a light blue body.

ctrl + b 또는 ctrl + shift + c 단축키로
선택한 요소에 해당하는 HTML 코드를 볼 수 있다.

서버의 Response와 HTML

```
<th scope="row">  
    <span>1 금</span>  
</th>  
<th scope="row" class="cal_sat">  
    <span class="pc_blue">2 토 </span>  
</th>
```

위의 형식에 맞는 문자열을 검색하기 위해서는

`<th scope="row"` 라는 문자열을 `r.text` 안에서 찾아야 한다.

String의 .find 메소드

```
1 s = "abcdabef"
2
3 print(s.find("ab"))
4
5 print(s.find("cdab"))
6
7 print(s.find("ab", 2))
8
9 print(s.find("ef", 0, 4))
10
11 # find는 인자를 세 개까지 전달받을 수 있다.
12 # 만약 해당 string을 찾지 못했을 경우에는 -1을 반환한다.
```

crawl.py

```
1 ...  
2  
3 r = requests.get(url + path + '?' + query)  
4  
5  
6 index = r.text.find('<th scope="row"')  
7  
8 print(r.text[index:index+38])  
9  
10  
11  
12
```

crawl.py

```
1 ...
2
3 r = requests.get(url + path + '?' + query)
4
5 index = r.text.find('<th scope="row"')
6
7 while index != -1:                                # 검색에 실패할 때까지
8     print(r.text[index:index+38])
9     next_index = r.text.find('<th scope="row"', index + 1)
10    index = next_index
11
12
```

crawl.py

```
1 ...
2
3 index = r.text.find('<th scope="row"')
4 while index != -1:                                # 검색에 실패할 때까지
5     final_index = r.text.find("</th>", index)
6     print(r.text[index:final_index])
7     next_index = r.text.find('<th scope="row"', index + 1)
8     title_start = r.text.find('title="', index, next_index)
9     if title_start != -1:
10         title_end = r.text.find('"', title_start + 7)
11         print(r.text[title_start+7:title_end])
12     index = next_index
```

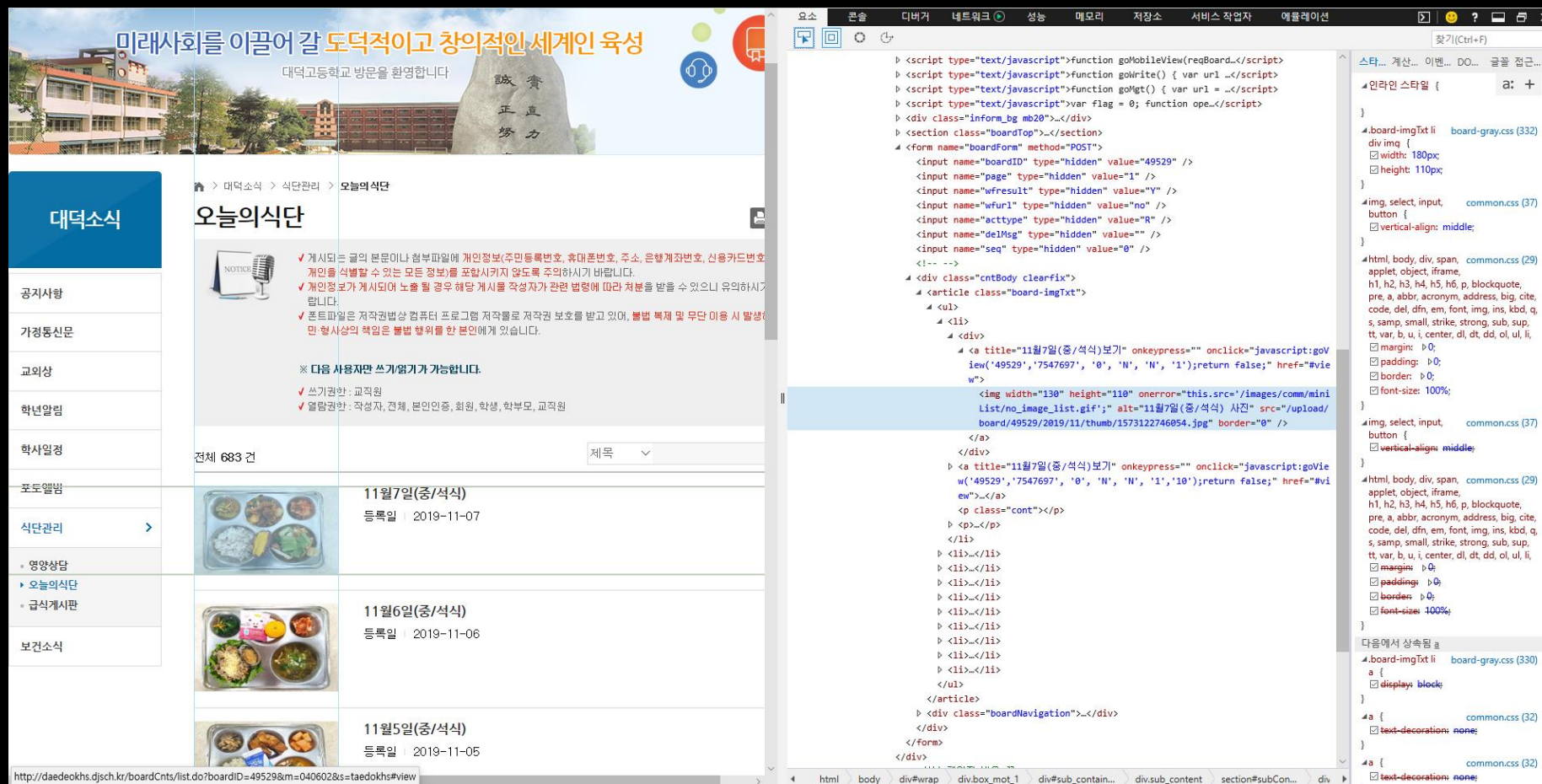
How to Parse?

```
<tr>
  <th scope="row"><span>1 금 </span></th>
  <td>
    <ul>
      <li>
        <a href="javascript:viewData('1075417');" title="학생연구의날">학생연구의날</a>
      </li>
    </ul>
  </td>
</tr>
<tr>
  <th scope="row" class="cal_sat"><span class="pc_blue">2 토 </span></th>
  <td class="cal_sat"></td>
</tr>
```

Diagram illustrating HTML parsing with annotations:

- index**: Points to the opening tag of the first row (`<tr>`).
- final_index**: Points to the closing tag of the first row (`</tr>`).
- title_start**: Points to the start of the `title` attribute in the first row's cell.
- title_start + 7**: Points to the start of the title text (`학생연구의날`).
- title_end**: Points to the end of the title text.
- next_index**: Points to the opening tag of the second row (`<tr>`).

식단 관리 페이지



The screenshot shows the KAIST Dining Management Page. The page title is "미래사회를 이끌어갈 도덕적이고 창의적인 세계인 육성" (Raising global leaders who are morally and creatively). The page is for "대덕소식" (Daedeok News) and "오늘의식단" (Today's Menu). The menu for November 7th (Wednesday) is displayed, showing a list of dishes and their prices. The browser developer tools are open, showing the HTML source code. The code includes a JavaScript function for mobile view, a form for board management, and a table for the menu. The table has columns for the date, day of the week, and the menu items. The menu items are listed in a table with columns for the date, day of the week, and the menu items. The table is titled "11월7일(중/석식)" and "등록일 2019-11-07". The table contains three rows of menu items, each with a photo and a description. The first row shows a bowl of rice with meat and vegetables. The second row shows a bowl of rice with meat and vegetables. The third row shows a bowl of rice with meat and vegetables. The table is titled "11월7일(중/석식)" and "등록일 2019-11-07".

ctrl + b 또는 ctrl + shift + c 단축키로
선택한 요소에 해당하는 HTML 코드를 볼 수 있다.

crawl.py

```
1 ...
2 url = 'http://daedeokhs.djsch.kr'
3 path = '/boardCnts/list.do'
4 query = 'boardID=49529&m=040602&s=taedokhs'
5
6 r = requests.get(url + path + '?' + query)
7
8 index = r.text.find('/upload/board/49529/')
9 while index != -1:
10     final_index = r.text.find('.jpg', index)
11     print(r.text[index:final_index+4])
12     index = r.text.find('/upload/board/49529/', index + 1)
```


식단 관리 페이지 (저화질)

위에서 crawl.py를 이용하여 수집한 path를 이용하여
<http://daedeokhs.djsch.kr/upload/board/49529/2019/11/thumb/1573122746054.jpg> 에 접속했을 때,



식단 관리 페이지 (고화질)

위에서 crawl.py를 이용하여 수집한 path를 수정하여
<http://daedeokhs.djsch.kr/upload/board/49529/2019/11/1573122746054.jpg> 에 접속했을 때,



crawl.py

```
1 ...
2 index = r.text.find('/upload/board/49529/')
3 while index != -1:
4     final_index = r.text.find('.jpg', index)
5     image_path = r.text[index:final_index+4]
6
7     image_path = image_path.replace('/thumb/', '/')
8     new_r = requests.get(url + image_path)
9     with open(str(index) + '.jpg', 'wb') as f:
10         f.write(new_r.content)
11
12     index = r.text.find('/upload/board/49529/', index + 1)
```

How to Parse?

```
<img src= index /upload/board/49529/2019/11/thumb/1573122746054.jpg'
onError="this.src='/images/comm/miniList/no_image_list.gif';"
width='130' height='110' border='0' alt="11월7일(중/석식) 사진"/>
final_index final_index + 4
```

예제로 놀아보기

- ① 공지사항, 가정통신문, 학년 알림 등의 글 제목을 수집하기
- ② 중식 이미지 뿐만 아니라, 석식 이미지도 수집하는 크롤러 만들기

BeautifulSoup4



- BeautifulSoup4란 라이브러리를 이용하면 훨씬 쉽게 HTML 문서에서 원하는 값들을 쉽게 뽑아올 수 있다.
- 관심 있는 사람들은 google에 검색

<https://twpower.github.io/84-how-to-use-beautiful-soup>

http://www.hanbit.co.kr/channel/category/category_view.html?cms_code=CMS4344562296