

파이썬을 이용한 데이터수집 및 스마트공장 견학

HTML/CSS
Crawling의 기초

2021년 1월 14일
안재관

금일 목표

- 간단한 HTML 문서 만들기
- 기본적인 HTML 태그 이해하기
- CSS의 역할 알아보기
- Selector를 이용해 원하는 HTML 요소들만 골라내기

웹의 3요소

- **HTML(Hyper Text Markup Language)**

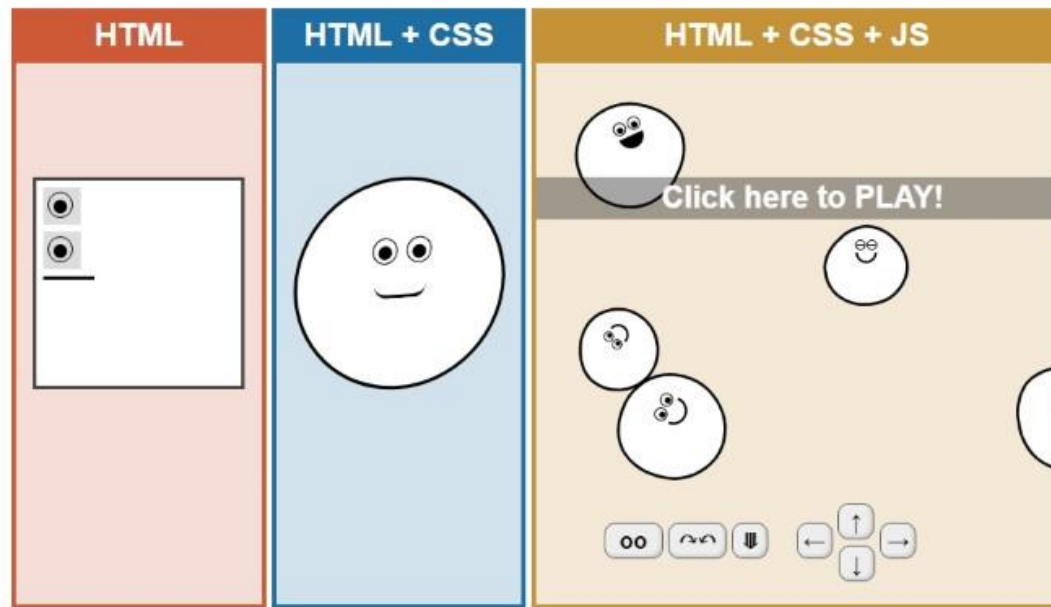
- 정적인 웹페이지

- **CSS(Cascading Style Sheet)**

- 웹페이지를 꾸미는 담당

- **Javascript**

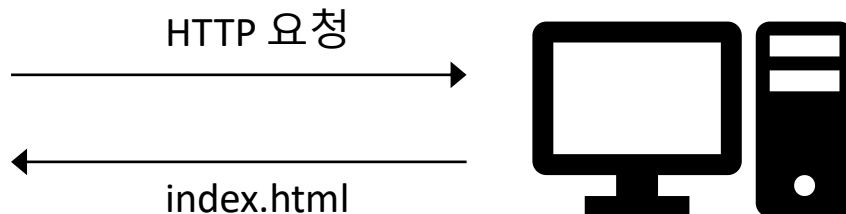
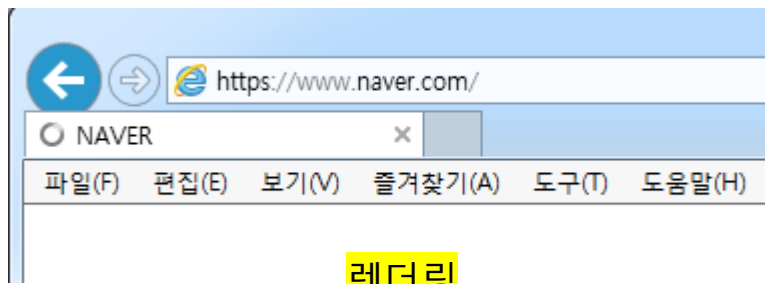
- 웹 페이지에서 동적인 부분 담당



웹 서버

● 웹 서버 (Web server)

- 웹 서버 (소프트웨어) : 웹 브라우저와 같은 클라이언트로부터 HTTP 요청을 받아들이고, HTML 문서와 같은 웹 페이지를 반환하는 컴퓨터 프로그램



웹 서버로부터 받은 index.html을 웹 브라우저가 해석한 후 렌더링해주는 화면을 우리가 보는 것.



HTML이란?

● HTML (Hyper Text Markup Language)

- 하이퍼 텍스트를 만드는 마크업 언어
- 하이퍼 텍스트는 쉽게 말해 링크가 있는 텍스트 문서
- 마크업 언어란 태그 (tag) 등을 이용하여 문서나 데이터의 조를 기술하는 언어

● HTML은 웹페이지의 뼈대/내용을 포함하며, 확장자는 *****.html, ***.htm**

● HTML 문서 실행 : 웹 브라우저

- 메모장을 실행
- 다음을 붙여넣기

```
<html>
<head><title> test </title></head>
<body>
hello
</body>
</html>
```

- test.html 저장 후 실행

HTML은 하이퍼텍스트 마크업 언어(HyperText Markup Language, 문화어: 초본문표식달기언어, 하이퍼본문표식달기언어)라는 의미의 웹 페이지를 위한 지배적인 마크업 언어다. HTML은 제목, 단락, 목록 등과 같은 본문을 위한 구조적 의미를 나타내는 것뿐만 아니라 링크, 인용과 그 밖의 항목으로 구조적 문서를 만들 수 있는 방법을

HTML이란?

● HTML 편집기

- Sublime Text 3 : 기능이 많은 텍스트 편집기
- <http://jsfiddle.net> : 편집과 실행을 동시에 가능케 함
 - 다음 내용을 JSFiddle의 HTML 칸에 붙여넣고 Run 누르기

```
<h1>테스트 문서입니다 크게</h1>
<h2>테스트 문서입니다 덜 크게</h2>
<h3>테스트 문서입니다 더 덜 크게</h3>
```

꺾쇠 사이에 있는 글을 태그라고 부릅니다.

태그를 사용해서 웹페이지를 시각화하고 기능을 부여 가능합니다.

글자를 진하게 만들수도 있고.

다른 웹사이트로 연결하는 링크를 만들수도 있으며.

그림을 보이게 할 수 있습니다.

태그는 중첩 가능하여 그림에 링크를 달수도 있습니다.



HTML 주요 태그

● HTML 태그는 시작과 끝이 있음

- <html> 내용 </html>, <head> 내용 </head>

주요 태그	태그 설명
<html>	HTML 문서임을 나타내는 태그
<head>	문서 정보, 메타 데이터, 외부 파일 정보 등을 기술
<body>	본문을 정의하는 태그로 이미지, 표, 문자를 표현
<p>	문단 구분을 위한 태그 (paragraph)
<table>	표를 정의하는 태그
 	줄바꿈 태그, 닫을 필요 없음 (line break)
<div>	HTML문서를 분할하는 태그; CSS에 많이 사용
	순서가 있는 목록을 표현하는 태그; 안에 사용
	순서가 없는 목록을 표현하는 태그; 안에 사용

- 태그들의 모든 정보는 <http://www.w3schools.com/tags/default.asp>

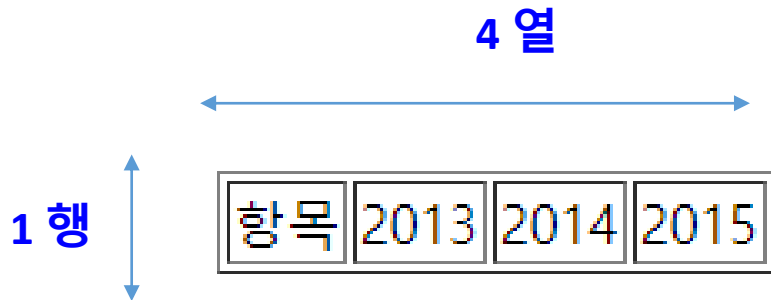
HTML 그 밖의 태그

- **img** : 그림을 웹페이지에 표시
 - ``
- **a** : 하이퍼링크
 - `링크할 내용`
- **p** : 문단 구분을 위한 태그 (paragraph)
 - `<p>문단 내용</p>`
- **<!-->** : 주석
 - `<!-- 나는 표시되지 않아요 -->`

HTML 표 만들기

● table, tr, td 태그

- table row (tr)
- table data (td)



```
<html>
<head><title> test </title></head>
<body>
<table border=1>
  <tr>
    <td> 항목 </td>
    <td> 2013 </td>
    <td> 2014 </td>
    <td> 2015</td>
  </tr>
</table>
</body>
</html>
```

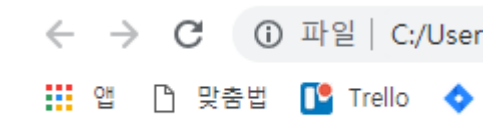
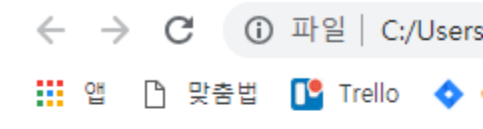
● 실습: 다음과 같은 표를 출력하여라.

항목	2013	2014	2015
매출액	100	200	300

HTML list 만들기

● ul, ol li 태그

- Unordered List (ul)
- Ordered List (ol)
- List content (li)

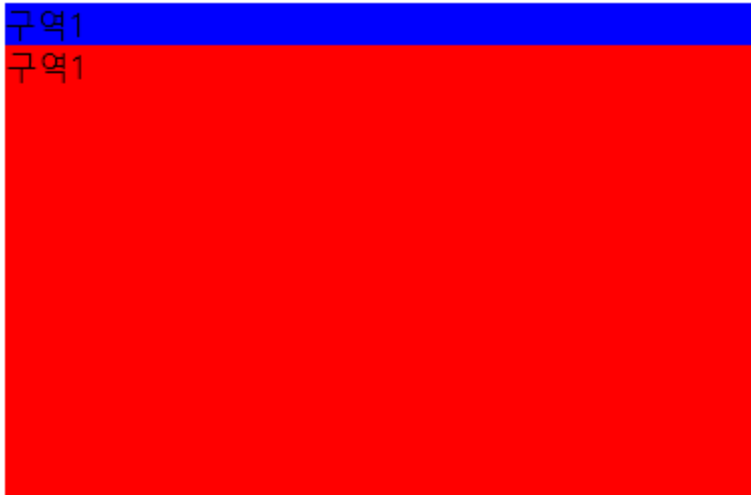


```
<html>
<head><title> pystock </title></head>
<body>
<ul>
  <li> HTML </li>
  <li> CSS </li>
  <li> Javascript </li>
</ul>
</body>
</html>
```

```
<html>
<head><title> pystock </title></head>
<body>
<ol>
  <li> HTML </li>
  <li> CSS </li>
  <li> Javascript </li>
</ol>
</body>
</html>
```

HTML 레이아웃 나누기

- **<div>** 태그는 Division 의 약자로 레이아웃을 나누는데 주로 사용



```
<html>
<head><title> pystock </title></head>
<body>
<div style="background-color:blue"> 구역1
</div>
<div style="background-color:red;
height:500px"> 구역1 </div>
</body>
</html>
```

CSS(Cascading Style Sheets) 소개

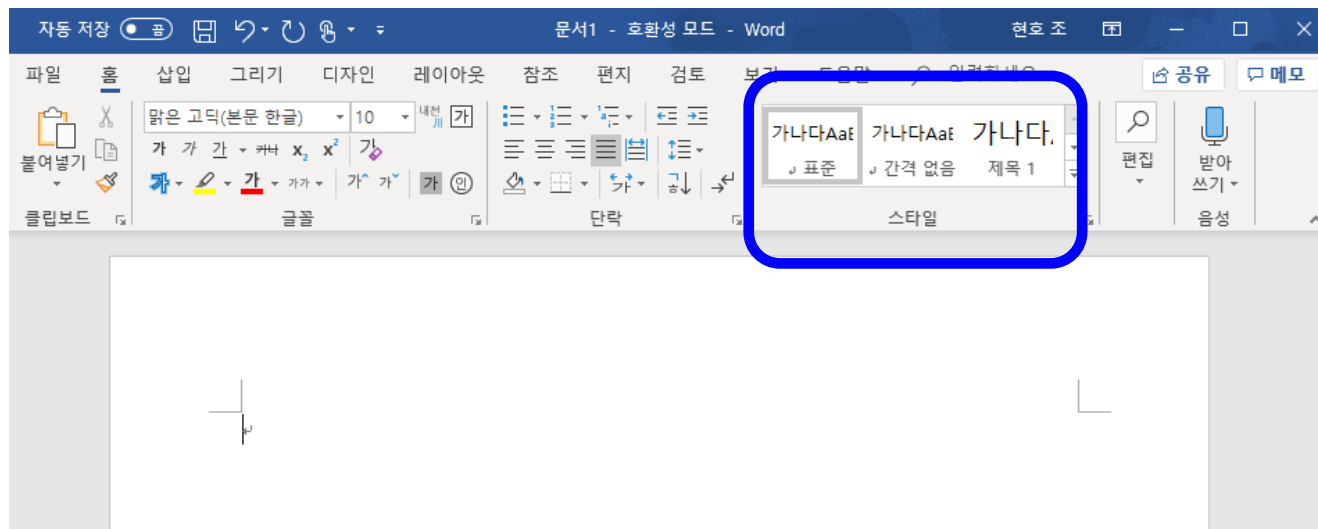
- CSS (Cascading Style Sheet)

- 마크업 언어가 실제 표시되는 방법을 기술하는 언어
- 문서를 꾸미는 역할을 담당

- 확장자는 *****.css**

- 마이크로소프트의 워드를 생각하면 쉬움

- 내용과 스타일을 분리



CSS(Cascading Style Sheets) 소개

- 특정 태그들에 특정 속성들을 일괄적으로 (cascading) 부여
- 사용
- 예시

```
selector {  
    property: value;  
}
```

- selector : 대상 태그들
- property : 디자인 속성명
- value : 디자인 값

```
p {  
    color: red;  
    background-color: blue;  
    font-family: Verdana;  
}
```

- <p> 태그들의 글자색은 red
- 배경색은 blue
- 글꼴은 Verdana

- 필요성 : HTML 내에 디자인을 위한 코드를 CSS로 분리 가능, 코드 양 ↓, 가독성 ↑
 - 모든 <p>에 <p style="color: red; background-color: blue;"></p> 대신 CSS로 한번에
 - 예쁜 CSS 있으면 나중에 재활용 가능, Python의 함수와 비슷함!
 - 웹사이트에 통일된 분위기 부여 가능

CSS Selector

- CSS는 태그에 스타일을 기술하는 용도로 사용되는 언어
- 기술한 스타일을 HTML의 어떤 태그에서 사용할지에 대해 정의하는 문법이 필요한데 이 문법을 selector라고 함
- CSS selector
 - Element selector: 해당되는 태그 전부에 스타일을 적용
 - Id selector: id 값을 가진 태그에만 스타일을 지정하는 방식 (문서에 유일)
 - Class selector: 같은 클래스의 이름을 갖는 모든 tag에 스타일 적용

CSS Selector

- **Element selector:** 해당되는 태그 전부에 스타일을 적용

```
<html>
<head>
  <style>
    p {color:blue}
  </style>
</head>
<body>
  <p> hello </p>
  <p> python </p>
</body>
</html>
```

hello
python

CSS Selector

- ID selector: id 값을 가진 태그에만 스타일을 지정하는 방식 (문서에 유일)

```
<html>
<head>
  <style>
    #title {color:blue; font-size:30}
  </style>
</head>
<body>
  <p id="title"> hello </p>
  <p> python </p>
</body>
</html>
```

hello

python

CSS Selector

- **Class selector:** 같은 클래스의 이름을 갖는 모든 tag에 스타일 적용

```
<html>
<head>
  <style>
    .title {color:blue; font-size:30}
  </style>
</head>
<body>
  <h1 class="title"> title-1 </h1>
  <h2> title-2 </h2>
  <h3 class="title"> title-3 </h3>
</body>
</html>
```

title-1

title-2

title-3

CSS Selector 예시

● 태그의 특수 속성 : id, class

- 태그의 형식 : <태그명 속성1="속성값1" 속성2="속성값2"> 텍스트 </태그명>
- 예시 HTML 코드

```
<산기대 class="시흥시 대학교">
  <학생 id="2019210000" class="에너지전기 경영" age="21">홍길동</학생>
  <학생 id="2018210001" class="에너지전기" age="22">김영희</학생>
  <학생 id="2019210002" class="경영 에너지전기 기계공학" age="21">김철수</학생>
  <교직원 id="B1354" class="경영 에너지전기" age="36">김재철</교직원>
  <교직원 id="AAAAA" class="에너지전기" age="99">이시영</교직원>
  <교직원 id="BBBBB" class="경영" age="99">허재석</교직원>
</산기대>
<서울대 class="대학교 서울시">
  <교직원 id="1111" class="경영" age="99">김교수</교직원>
  <학생 id="2222" class="산업공학" age="36">양양양</학생>
</서울대>
```

- id와 class 속성들은 selector에서 태그들을 더 자세하게 선택될 수 있게 함
 - id : 문서 내에서 고유해야함, #로 선택
 - class : .로 선택, and로 선택하고자 할 때는 붙여서 .속성값1.속성값2, 순서 무관

CSS Selector

● Selector : HTML 문서 내에서 찾고자 하는 태그들을 골라내는 역할

```
<산기대 class="시흥시 대학교">
  <학생 id="2019210000" class="에너지전기 경영" age="21">홍길동</학생>
  <학생 id="2018210001" class="에너지전기" age="22">김영희</학생>
  <학생 id="2019210002" class="경영 에너지전기 기계공학" age="21">김철수</학생>
  <교직원 id="B1354" class="경영 에너지전기" age="36">김재철</교직원>
  <교직원 id="AAAAA" class="에너지전기" age="99">이시영</교직원>
  <교직원 id="BBBBB" class="경영" age="99">허재석</교직원>
</산기대>
<서울대 class="대학교 서울시">
  <교직원 id="1111" class="경영" age="99">김교수</교직원>
  <학생 id="2222" class="산업공학" age="36">양양양</학생>
</서울대>
```

- A : A라는 태그명들을 선택
 - 학생 : 홍길동, 김영희, 김철수, 양양양

CSS Selector

● Selector : HTML 문서 내에서 찾고자 하는 태그들을 골라내는 역할

```
<산기대 class="시흥시 대학교">
  <학생 id="2019210000" class="에너지전기 경영" age="21">홍길동</학생>
  <학생 id="2018210001" class="에너지전기" age="22">김영희</학생>
  <학생 id="2019210002" class="경영 에너지전기 기계공학" age="21">김철수</학생>
  <교직원 id="B1354" class="경영 에너지전기" age="36">김재철</교직원>
  <교직원 id="AAAAA" class="에너지전기" age="99">이시영</교직원>
  <교직원 id="BBBBB" class="경영" age="99">허재석</교직원>
</산기대>
<서울대 class="대학교 서울시">
  <교직원 id="1111" class="경영" age="99">김교수</교직원>
  <학생 id="2222" class="산업공학" age="36">양양양</학생>
</서울대>
```

- A : A라는 태그명들을 선택
 - 서울대

CSS Selector

● Selector : HTML 문서 내에서 찾고자 하는 태그들을 골라내는 역할

```
<산기대 class="시흥시 대학교">
  <학생 id="2019210000" class="에너지전기 경영" age="21">홍길동</학생>
  <학생 id="2018210001" class="에너지전기" age="22">김영희</학생>
  <학생 id="2019210002" class="경영 에너지전기 기계공학" age="21">김철수</학생>
  <교직원 id="B1354" class="경영 에너지전기" age="36">김재철</교직원>
  <교직원 id="AAAAA" class="에너지전기" age="99">이시영</교직원>
  <교직원 id="BBBBB" class="경영" age="99">허재석</교직원>
</산기대>
<서울대 class="대학교 서울시">
  <교직원 id="1111" class="경영" age="99">김교수</교직원>
  <학생 id="2222" class="산업공학" age="36">양양양</학생>
</서울대>
```

- #A : id의 속성값이 A인 태그를 선택
 - #B1354 : 김재철

CSS Selector

● Selector : HTML 문서 내에서 찾고자 하는 태그들을 골라내는 역할

```
<산기대 class="시흥시 대학교">
  <학생 id="2019210000" class="에너지전기 경영" age="21">홍길동</학생>
  <학생 id="2018210001" class="에너지전기" age="22">김영희</학생>
  <학생 id="2019210002" class="경영 에너지전기 기계공학" age="21">김철수</학생>
  <교직원 id="B1354" class="경영 에너지전기" age="36">김재철</교직원>
  <교직원 id="AAAAA" class="에너지전기" age="99">이시영</교직원>
  <교직원 id="BBBBB" class="경영" age="99">허재석</교직원>
</산기대>
<서울대 class="대학교 서울시">
  <교직원 id="1111" class="경영" age="99">김교수</교직원>
  <학생 id="2222" class="산업공학" age="36">양양양</학생>
</서울대>
```

- .A : class의 속성값 중에 A가 있는 태그들을 선택
 - .경영 : 홍길동, 김철수, 양용석, 허재석, 김교수

CSS Selector

● Selector : HTML 문서 내에서 찾고자 하는 태그들을 골라내는 역할

```
<산기대 class="시흥시 대학교">
  <학생 id="2019210000" class="에너지전기 경영" age="21">홍길동</학생>
  <학생 id="2018210001" class="에너지전기" age="22">김영희</학생>
  <학생 id="2019210002" class="경영 에너지전기 기계공학" age="21">김철수</학생>
  <교직원 id="B1354" class="경영 에너지전기" age="36">김재철</교직원>
  <교직원 id="AAAAA" class="에너지전기" age="99">이시영</교직원>
  <교직원 id="BBBBB" class="경영" age="99">허재석</교직원>
</산기대>
<서울대 class="대학교 서울시">
  <교직원 id="1111" class="경영" age="99">김교수</교직원>
  <학생 id="2222" class="산업공학" age="36">양양양</학생>
</서울대>
```

- .A.B : class의 속성값 중에 A와 B를 동시에 갖고 있는 태그들을 선택
 - .경영.에너지전기 : 홍길동, 김철수, 김재철
 - .에너지전기.경영 : 같은 결과, 순서 무관

CSS Selector

● Selector : HTML 문서 내에서 찾고자 하는 태그들을 골라내는 역할

```
<산기대 class="시흥시 대학교">
  <학생 id="2019210000" class="에너지전기 경영" age="21">홍길동</학생>
  <학생 id="2018210001" class="에너지전기" age="22">김영희</학생>
  <학생 id="2019210002" class="경영 에너지전기 기계공학" age="21">김철수</학생>
  <교직원 id="B1354" class="경영 에너지전기" age="36">김재철</교직원>
  <교직원 id="AAAAA" class="에너지전기" age="99">이시영</교직원>
  <교직원 id="BBBBB" class="경영" age="99">허재석</교직원>
</산기대>
<서울대 class="대학교 서울시">
  <교직원 id="1111" class="경영" age="99">김교수</교직원>
  <학생 id="2222" class="산업공학" age="36">양양양</학생>
</서울대>
```

- aaaa bbbb : aaaa에 해당하는 태그들 속에 있는 bbbb에 해당하는 태그들을 선택
 - .시흥시 학생 : 홍길동, 김영희, 김철수

CSS Selector

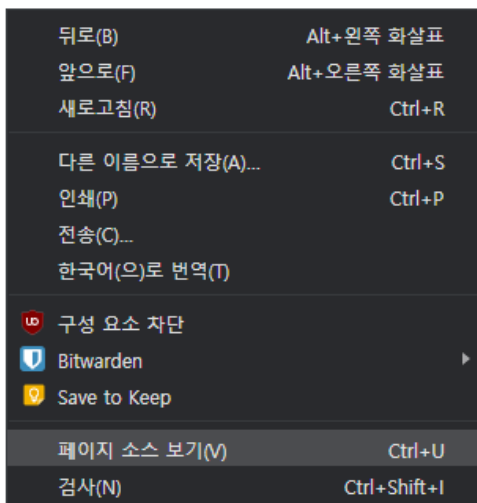
● Selector : HTML 문서 내에서 찾고자 하는 태그들을 골라내는 역할

```
<산기대 class="시흥시 대학교">
  <학생 id="2019210000" class="에너지전기 경영" age="21">홍길동</학생>
  <학생 id="2018210001" class="에너지전기" age="22">김영희</학생>
  <학생 id="2019210002" class="경영 에너지전기 기계공학" age="21">김철수</학생>
  <교직원 id="B1354" class="경영 에너지전기" age="36">김재철</교직원>
  <교직원 id="AAAAA" class="에너지전기" age="99">이시영</교직원>
  <교직원 id="BBBBB" class="경영" age="99">허재석</교직원>
</산기대>
<서울대 class="대학교 서울시">
  <교직원 id="1111" class="경영" age="99">김교수</교직원>
  <학생 id="2222" class="산업공학" age="36">양양양</학생>
</서울대>
```

- aaaabbbb : aaaa와 bbbb를 동시에 만족하는 태그들을 선택
 - 교직원.경영 : 김재철, 허재석, 김교수

HTML 소스 분석

- (우리의) 목표 : 웹페이지 내에서 원하는 정보를 찾기 위함
- 소스 보기 (Chrome 기준, 타 브라우저도 유사)
 - 소스를 보고자 하는 웹페이지에서 우클릭하여 “페이지 소스 보기” 클릭



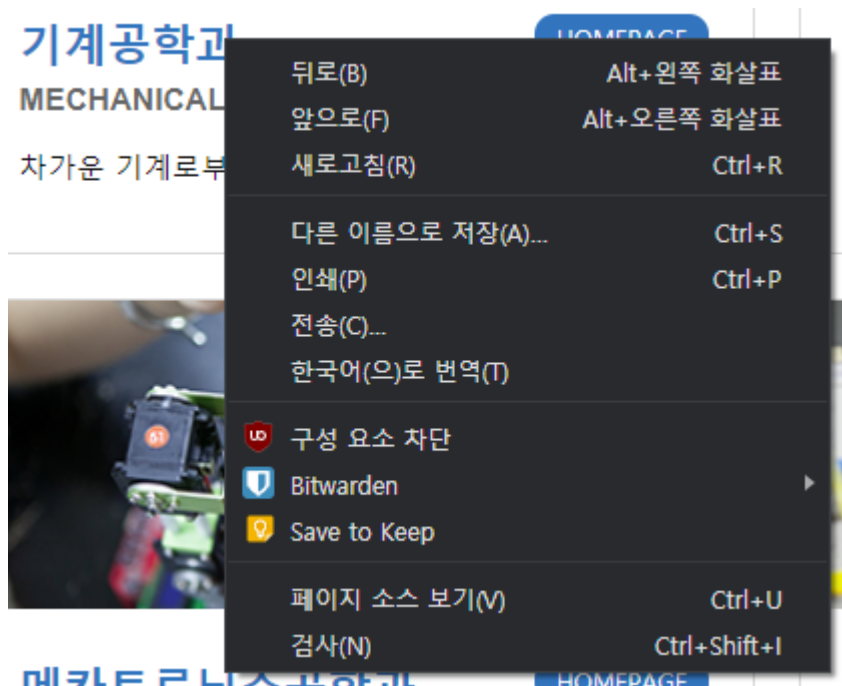
```
<!DOCTYPE HTML>
<html lang="ko">
<head>
  <meta charset="euc-kr">
  <meta http-equiv="X-UA-Compatible" content="IE=edge, chrome=1">
  <meta name="format-detection" content="telephone=no">
  <title>한국산업기술대학교 - 대학</title>
  <!--[if !IE 8]><!--><link rel="stylesheet" type="text/css" href="/front/styles/font.css"><!--<!-->
  <link media="all" rel="stylesheet" type="text/css" href="/front/addons/font-awesome/css/font-awes
  <link media="all" rel="stylesheet" type="text/css" href="/front/plugins/jquery-ui.min.css">
  <link media="all" rel="stylesheet" type="text/css" href="/front/plugins/jquery-ui.theme.css">
  <link media="all" rel="stylesheet" type="text/css" href="/front/styles/common.css">
  <link media="all" rel="stylesheet" type="text/css" href="/front/styles/default.css">
  <link media="all" rel="stylesheet" type="text/css" href="/front/styles/device.css">
  <!--[if IE 8]><link media="all" rel="stylesheet" type="text/css" href="styles/conditional/ie8.css
  <!--[if IE 9]><link media="all" rel="stylesheet" type="text/css" href="styles/conditional/ie9.css
  <script type="text/javascript" src="/front/plugins/jquery.1.11.1.min.js"></script>
  <script type="text/javascript" src="/front/plugins/jquery-ui.1.11.2.min.js"></script>
  <script type="text/javascript" src="/front/plugins/jquery-migrate.1.2.1.min.js"></script>
  <script type="text/javascript" src="/front/plugins/jquery.hoverintent.1.8.1.min.js"></script>
  <script type="text/javascript" src="/front/plugins/jquery.cookie.1.4.1.min.js"></script>
  <script type="text/javascript" src="/front/scripts/default.js"></script>
  <script type="text/javascript" src="/js/common.js"></script>
  <script type="text/javascript" src="/js/resize.js"></script>
  <script type="text/javascript" src="/js/cms/main/cor/main.js?20190522155732"></script>
  <script type="text/javascript">
  <!--
  var menuCodeValue;
  var menuCodeName;
  menuCodeValue = "003001";
  menuCodeName = "대학/대학원";
```

- 눈에 쉽게 들어오지 않고 원하는 내용을 찾기 어려움

HTML 소스 분석

● 검사/요소 검사

- 웹페이지에서 관심이 있는 항목에서 우클릭, 검사 클릭
- 예시
 - 한국산업기술대학교 학과 홈페이지에서 학과들의 소스가 어떻게 되어있는지 보려면, 학과 하나에서 "우클릭 - 검사"
 - <http://www.kpu.ac.kr/contents/main/cor/kcollege.html>



HTML 소스 분석

● 검사/요소 검사

- 우클릭을 한 곳의 소스가 구체적으로 나옴



기계공학과

MECHANICAL ENGINEERING

차가운 기계로부터 뜨거운 젊음의 열정을 배운다

HOMEPAGE

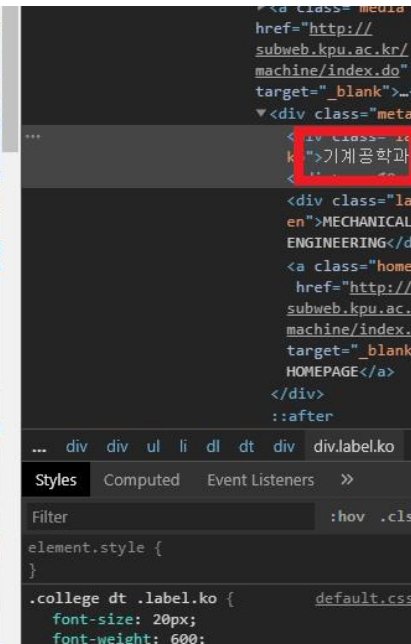


기계설계공학과

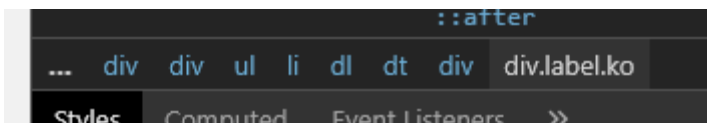
MECHANICAL DESIGN ENGINEERING

세상을 움직이는 새로운 기계를 창조한다

HOMEPAGE



- 원하는 내용이 어떤 태그들 아래에 속해 있는지 정리되어 있음



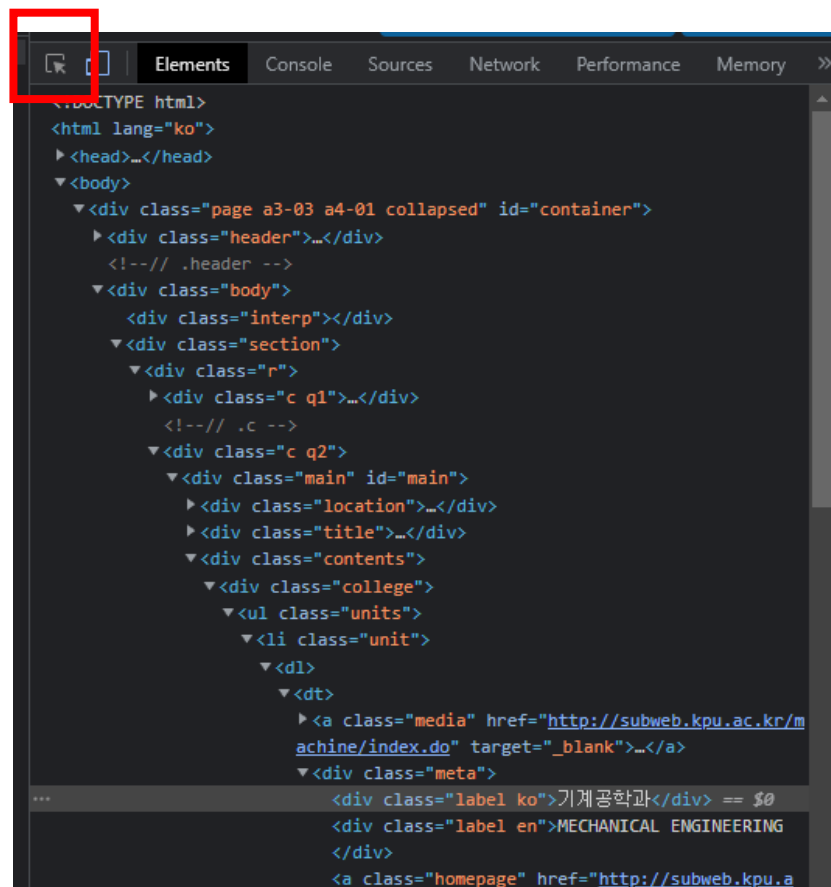
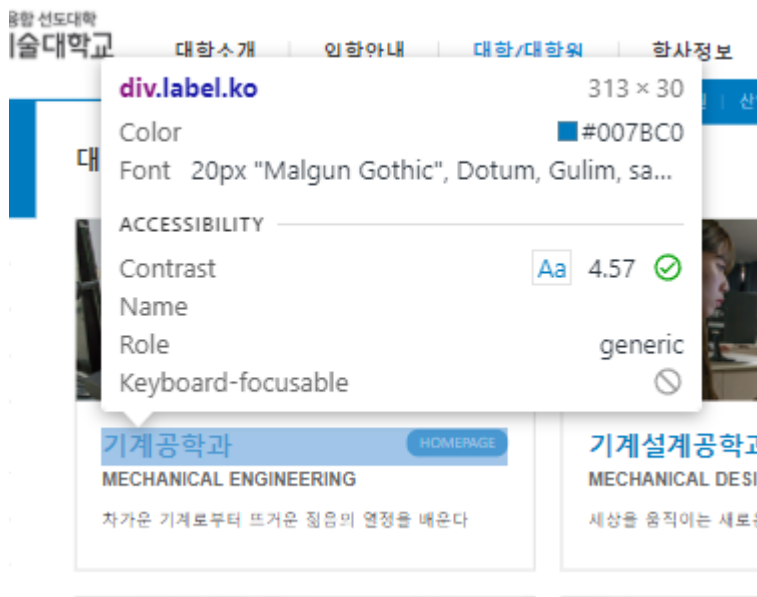
- 소스에 마우스를 올리면 웹페이지에서 어느 부분인지 표시해줌



HTML 소스 분석

● 검사/요소 검사

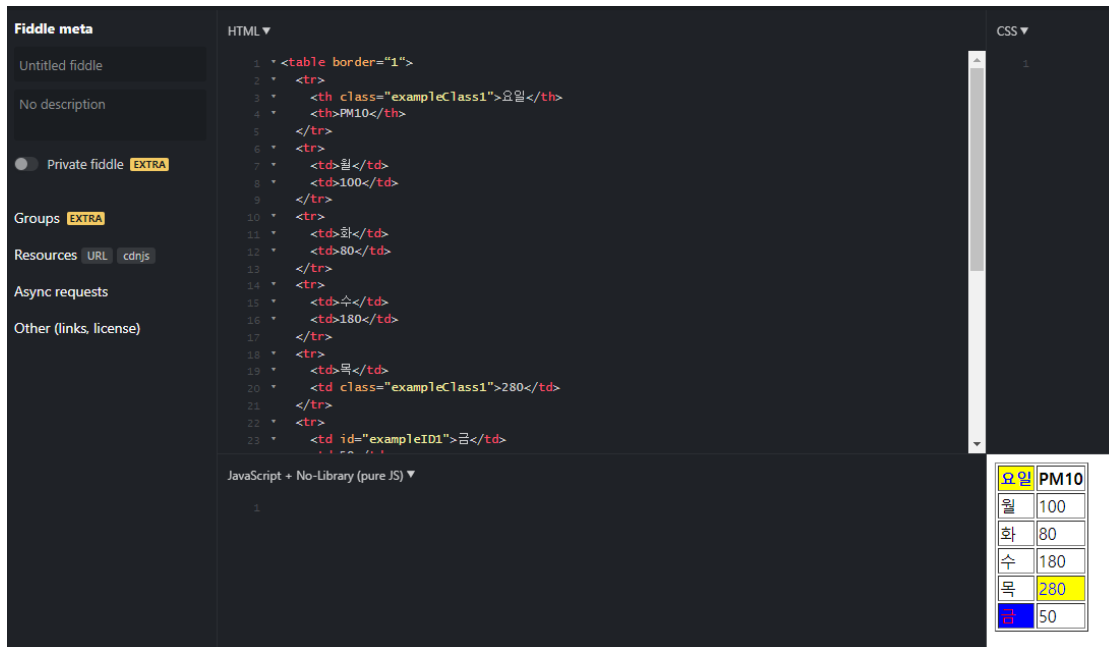
- 반대로 저 네모마우스 모양을 클릭하면 웹페이지에 특정 부분의 소스를 제공



HTML 소스 분석

● Selector 실습#1

- <http://147.46.178.16:33333/table.html> 들어가서 소스보기
- <https://jsfiddle.net/> 들어가기
- HTML 부분에 소스에서 복사해서 넣기
- Run 누르기



The screenshot shows the jsfiddle.net interface with the following content:

Fiddle meta

- Untitled fiddle
- No description
- Private fiddle **EXTRA**
- Groups **EXTRA**
- Resources **URL** **cdnjs**
- Async requests
- Other (links, license)

HTML

```
1 <table border="1">
2 <tr>
3 <th class="exampleClass">요일</th>
4 <th>PM10</th>
5 </tr>
6 <tr>
7 <td>월</td>
8 <td>100</td>
9 </tr>
10 <tr>
11 <td>화</td>
12 <td>80</td>
13 </tr>
14 <tr>
15 <td>수</td>
16 <td>180</td>
17 </tr>
18 <tr>
19 <td>목</td>
20 <td class="exampleClass">280</td>
21 </tr>
22 <tr>
23 <td id="exampleID1">금</td>
```

CSS

```
1
```

JavaScript + No-Library (pure JS)

```
1
```

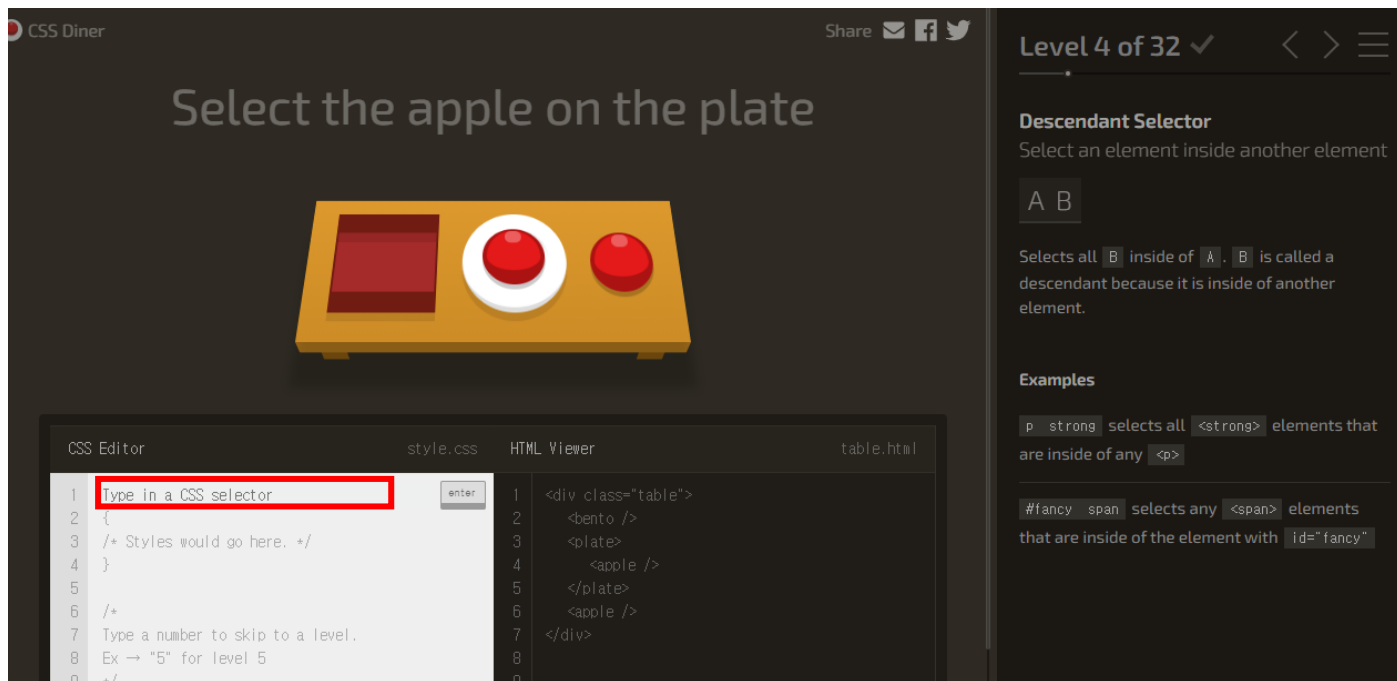
Preview

요일	PM10
월	100
화	80
수	180
목	280
금	50

HTML 소스 분석

● Selector 실습#2

- <http://flukeout.github.io/>
- 가상의 태그명들을 사용한 퀴즈들
- 문제에 맞는 물품들만을 골라낼 수 있는 selector를 적어야함
- 몇 개는 사이트 작성자가 틀리게 한 것도 있기에 1~11번 문제만 풀어보기
- 빨간 네모에 selector를 적고 enter 누르기



● Selector 실습#3

- <http://www.kpu.ac.kr/contents/main/cor/kcollege.html> 에서 Ctrl+F로 다음 정보들 선택
 - 한글 학과명 : .label.ko
 - 영문 학과명 : .label.en
 - 학과 설명 : .desc
- 간단한 홈페이지에서는 위와 같이 찾으면 웬만하면 끝남
- 복잡한 홈페이지에서는 해당 class를 재사용할 가능성이 있음
- 따라서 더 구체적으로 요소를 저장할 필요가 있음
- 위 정보들을 다른 selector들을 사용하여 다양한 방식으로 검색
 - 한글 학과명을 선택하는 다양한 방법
 - .meta .ko : meta class 아래의 ko class
 - li dl dt div div.ko : li 태그 아래 dl 태그 아래 dt 태그 아래 div 태그 아래 div태그 중에서 ko를 class로 갖는 태그
 -