Data Crawling

데이터 크롤링

류영표 강사
youngpyoryu@dongguk.edu

강사 소개

류영표
Youngpyo Ryu

동국대학교 수학과/응용수학 석사수료

現 Upstage AI X 네이버 부스트 캠프 1기,2기 멘토

現 Innovation on Quantum & CT(IQCT) 이사

前 한국파스퇴르연구소 Image Mining 인턴(Deep learning)

前 ㈜셈웨어(수학컨텐츠, 데이터 분석 개발 및 연구인턴)

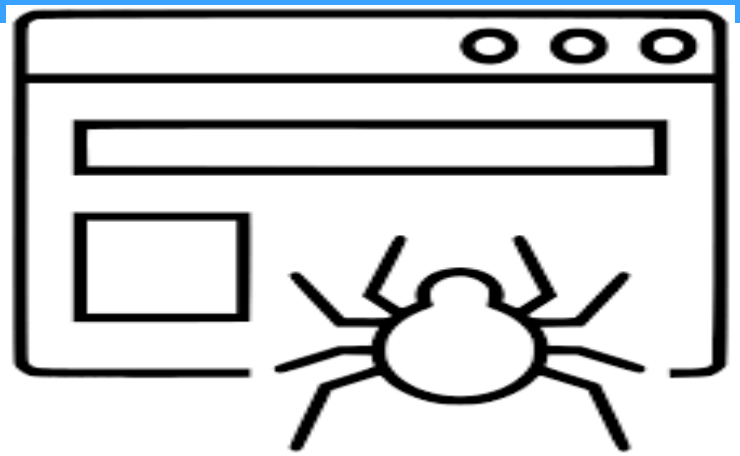
강의 경력

- 현대자동차 연구원 강의 (인공지능/머신러닝/딥러닝/강화학습)
- 딥러닝 집중 교육과정 강사
- (재)윌튼블록체인 6일 과정 (파이썬기초, 크롤링,머신러닝)
- ㈜ 모두의연구소 Aiffel 1기 퍼실리테이터(인공지능 교육)
- 인공지능 자연어처리(NLP) 기업데이터 분석 전문가 양성과정 멘토
- 고려대학교 선도대학 소속 30명 딥러닝 집중 강의
- 메가 IT 아카데미(파이썬, 빅데이터 강사)
- 이젠 종로 아카데미(파이썬, ADSP 강사)
- 최적화된 도구(R/파이썬)을 활용한 애널리스트 양성과정(국비과정) 강사
- 한화, 하나금융사 교육
- 인공지능 신뢰성 확보를 위한 실무 전문가 자문위원

주요 프로젝트 및 기타사항

- 제1회 인공지능(AI)기반 데이터사이언티스트 전문가 양성과정 최우수상 수상(Q&A 챗봇)
- 인공지능(AI)기반 데이터사이언티스트 전문가 양성과정 1기 수료
- 제 1회 산업 수학 스터디 그룹 (질병에 영향을 미치는 유전자 정보 분석)
- 제 4,5회 산업 수학 스터디 그룹 (피부암, 유방암 분류)
- 빅데이터 여름학교 참석 (혼잡도를 최소화하는 새로운 노선 건설 위치의 최적화 문제)

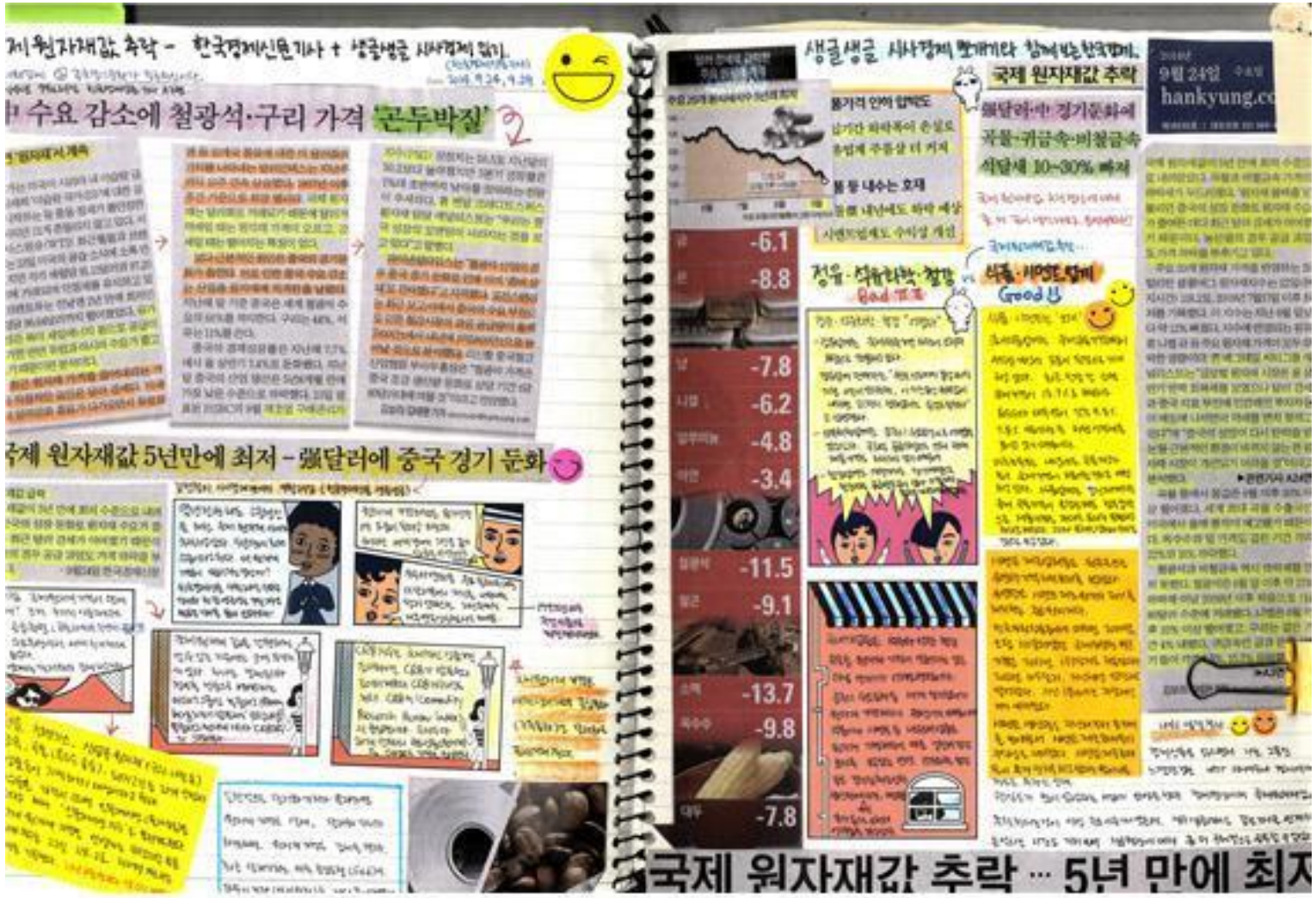
1. Data crawling



- 사전적으로 기어다니는 것을 뜻함.
- Web상을 돌아다니면서 수집하는 행위
- 주로 인터넷 상의 웹페이지(html, 문서 등)를 수집해서 분류하고 저장하는 것을 의미함.
- 데이터 수집보다는 여러 웹페이지를 돌아다닌다는 뜻이 강하며, 데이터가 어디에 저장되어 있는지 위치에 대한 분류 작업이 크롤링의 주요 목적.

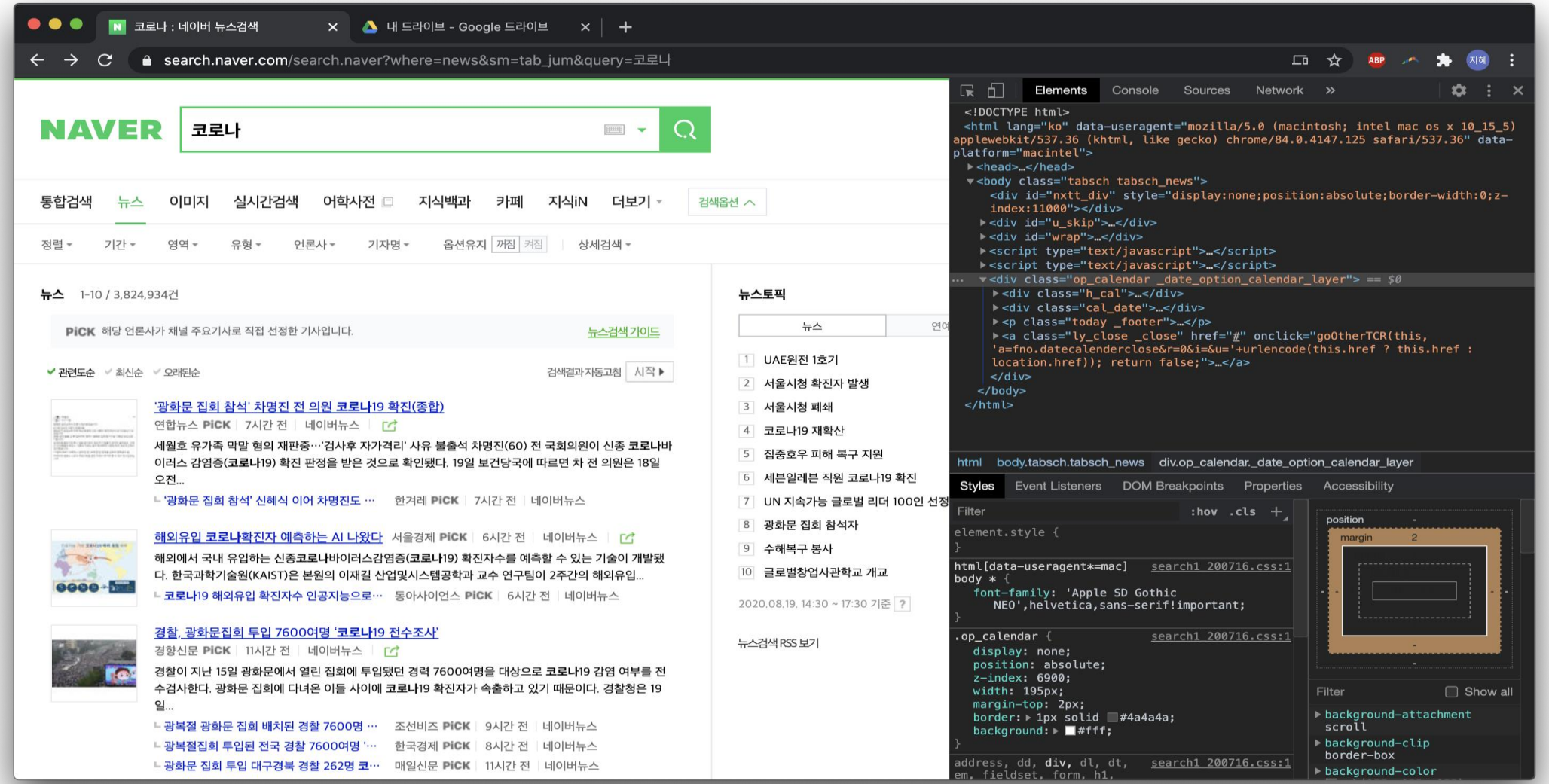
1. Scrap

1) 응용 분야



출처 : https://brunch.co.kr/@kpf10/290

Web Crawling



- 특정한 웹 페이지 내용에 원하는 부분을 내가 원하는 형식으로 만드는 것이다.

출처 : https://velog.io/@new_wisdom/%EB%A9%8B%EC%82%AC-%EB%B0%A9%ED%95%99-7-%EC%9B%B9-%ED%81%AC%EB%A1%A4%EB%A7%81Web-Crawling

Web Crawling

- 크롤링(Crawling) : 웹 크롤러(Crawler)라는 단어에서 시작된 말.

  크롤러란 조직적, 자동화된 방법으로 월드와이드 웹을 탐색하는 컴퓨터

  프로그램(출처 : 위키백과)

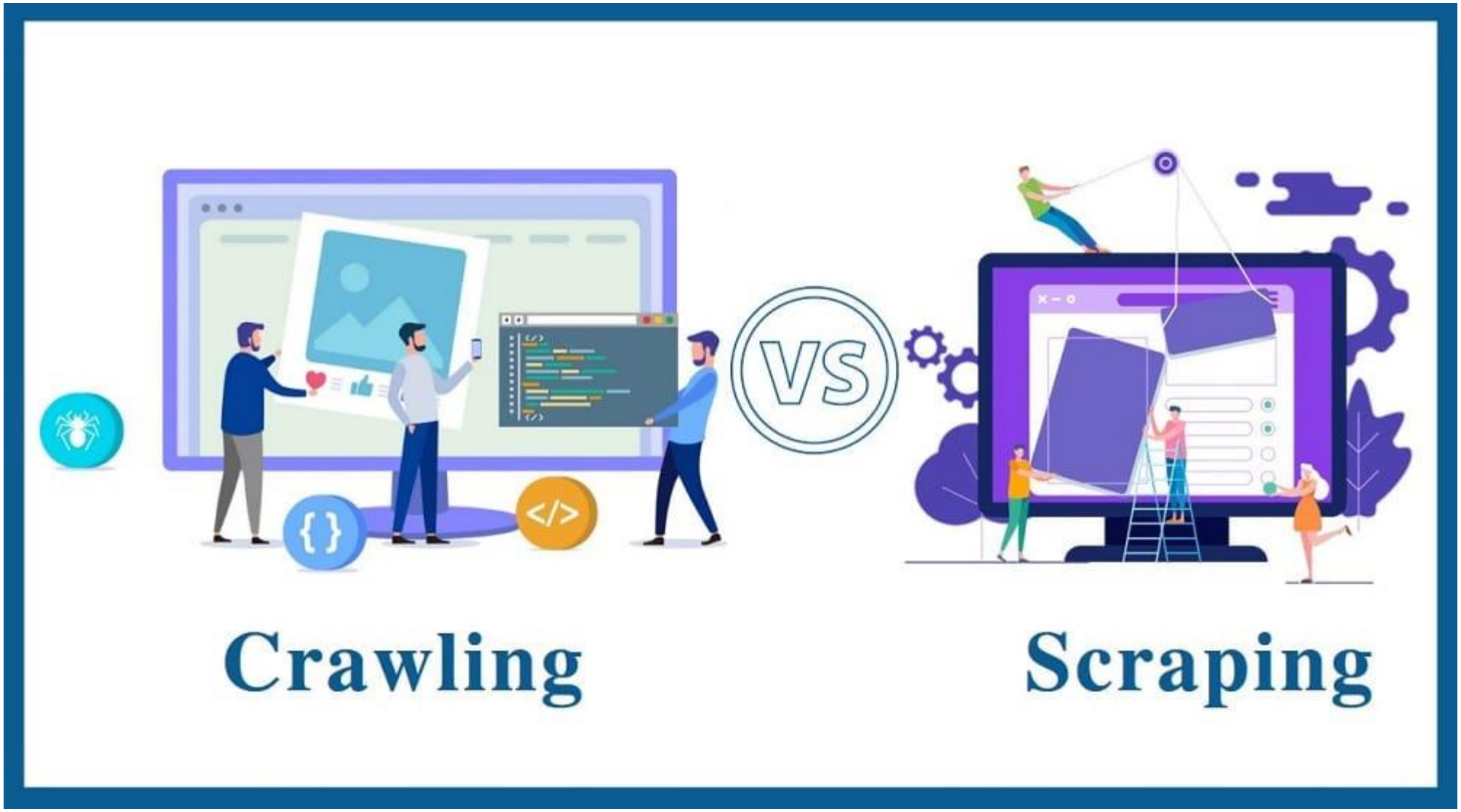
- 파싱(Parsing) : 어떤 페이지(문서, html)에서 내가 원하는 데이터를

  특정 패턴이나 순서로 추출하여 정보를 가공하는 것

- 스크래핑(Scraping) : HTTP을 통해 웹 사이트의 내용을 긁어다 원하

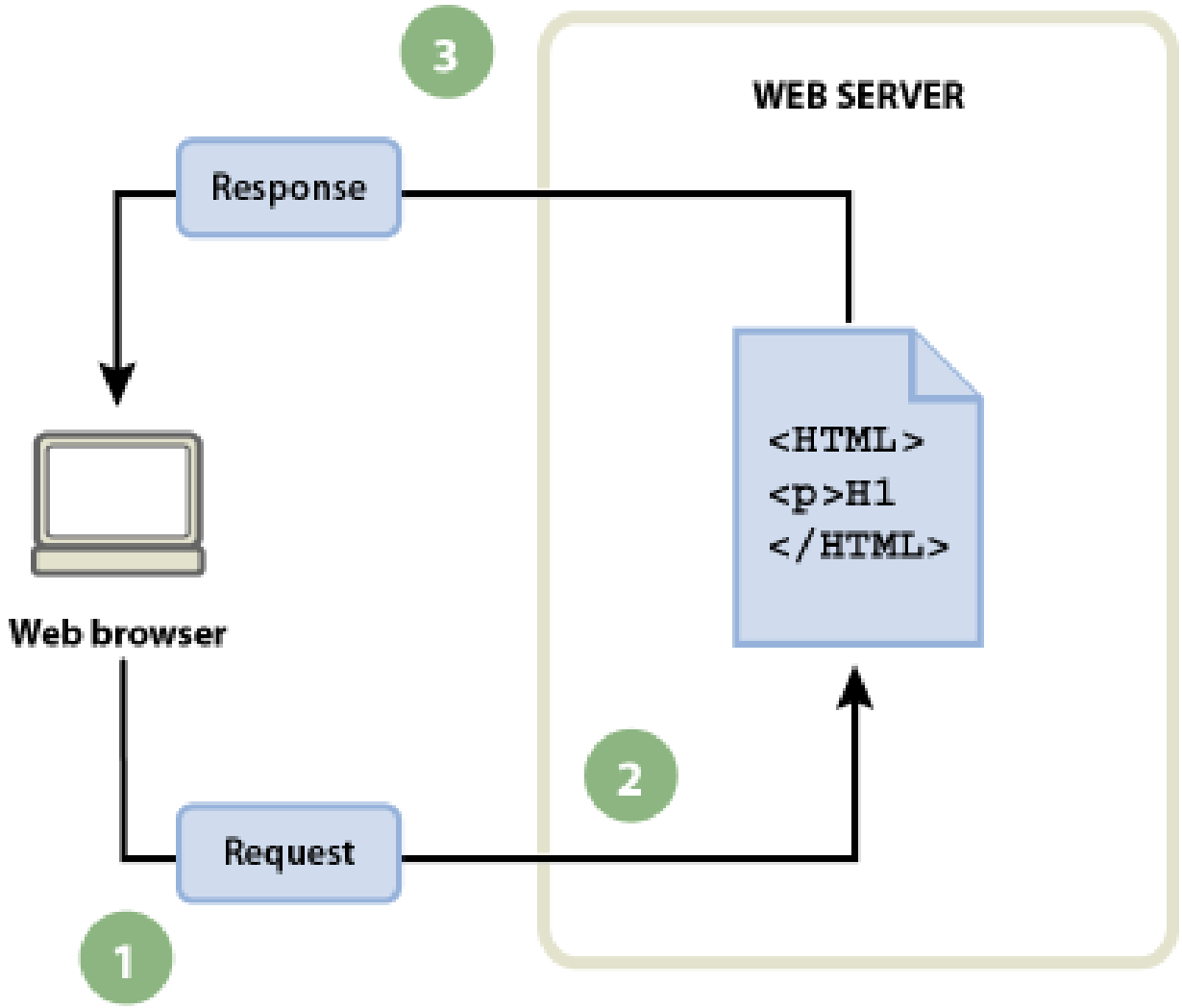
  는 형태로 가공 하는 것. 크롤링도 일종의 스크래핑의 기술.

Data Crawling VS Scraping



출처 : https://www.bestproxyreviews.com/crawling-vs-scraping/

Web page



출처 : https://getput.tistory.com/1

HTML, CSS, JavaScript

정보 및 설계도

디자인 및 스타일링

기능과 효과

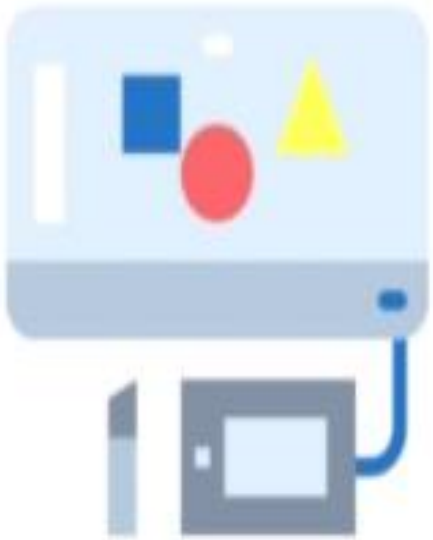
웹 사이트 제작 = 건물 짓기

· HTML

건물 설계도

· CSS

인테리어 디자인

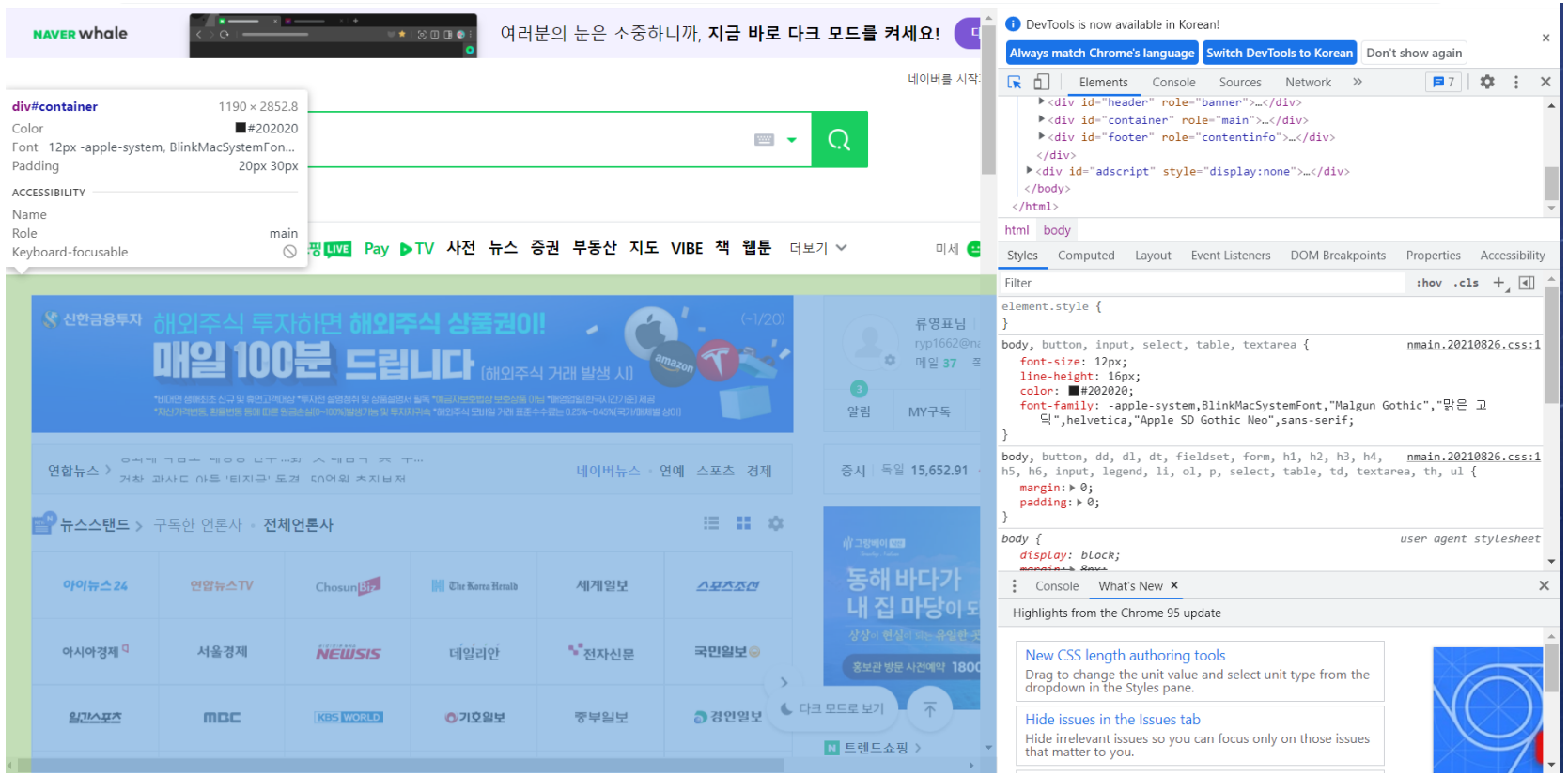


· JavaScript

기능과 효과

HTML

Hyper Text Markup Language

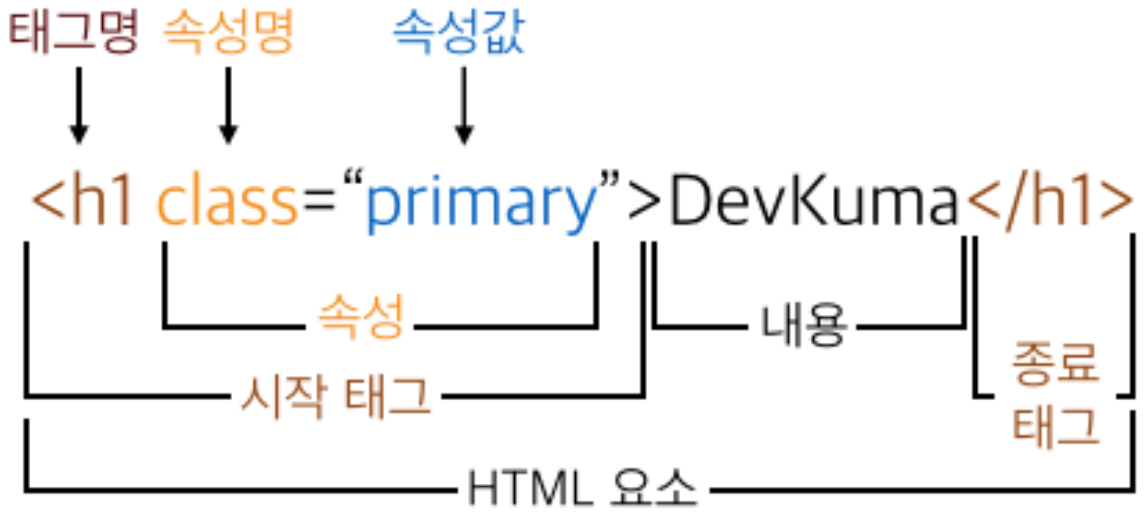


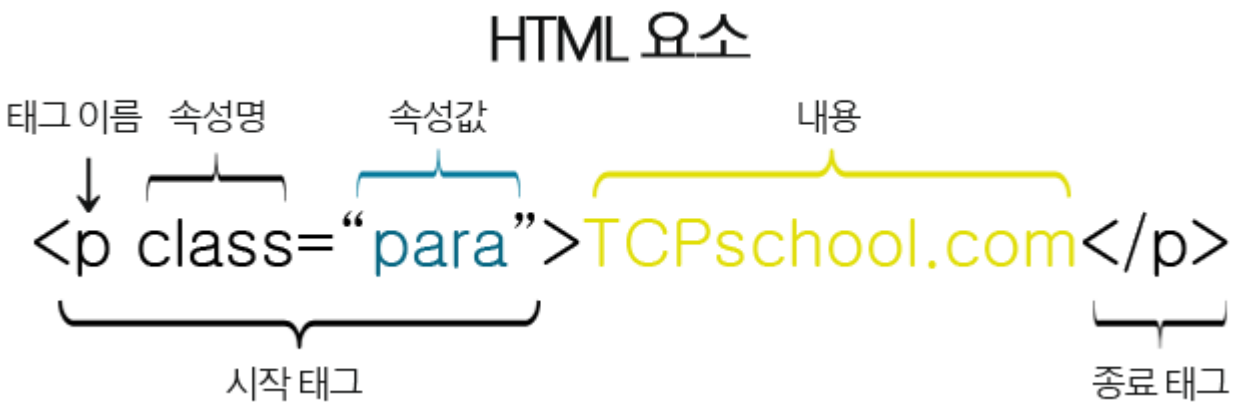
웹페이지의 내용과 구조를 담당하는 언어

HTML

출처 : https://lh3.googleusercontent.com/proxy/oT1nc8ErGIBEXjABwCI948XBBaWI-34KUJF7XsR17TwTGog5QCIHDwpeNLLqt8MaKpbmeaJLdVb5sA7bSYjBQUgg_5hAW-Im2KRPbkFQ

HTML





출처 : http://tcpschool.com/html/html_intro_elementStructure

HTML

```
<!DOCTYPE html>
<html>
    <head>
        <title>My Homepage</title>
    </head>
    <body>
        Welcome to my homepage!
    </body>
</html>
```

<!DOCTYPE html>

- HTML 문서임을 선언

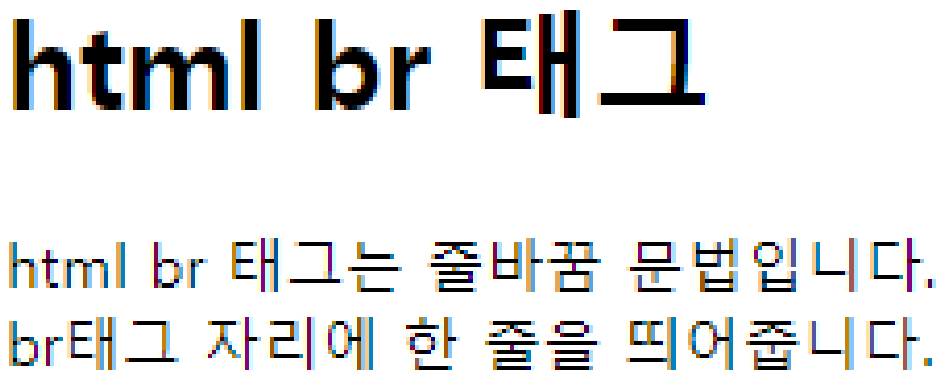
<html> … </html>

- HTML 문서의 시작과 끝을 의미
- 다른 요소들은 <html> 요소 안에 입력

HTML Tag -

줄 바꿈을 하려면 직접 줄 바꿈을 한다는 명령을 적어 주어야 하며, HTML 에서는
를 통하여 줄바꿈을 함.

```
<!DOCTYPE html>
<html>
  <head></head>
  <body>
    <h1>html br 태그</h1>
  html br 태그는 줄바꿈 문법입니다.
    <br>
  br태그 자리에 한 줄을 띄어줍니다.
  </body>
</html>
```



HTML Tag - …

Paragraph의 약자로, 문단을 나타냄.

문단 간에는 한 줄의 간격 생성

```
<html>
  <body>
    <p>first paragraph</p>
    <p>second paragraph</p>
    <p>
      new line<br>
      third paragraph
    </p>
  </body>
</html>
```

first paragraph

second paragraph

new line
third paragraph

HTML Tag - <b>

Blod의 약자

 태그도 같은 기능을 하며, 최신 표준은 <b> 태그보다는 태그를 권고 함.

```
<html>
<body>
  <b>bold content</b>
  normal content
</body>
</html>
```

bold content normal content

HTML Tag - <h#>

섹션, 문단의 제목을 나타내며 숫자가 작을수록 글자의 크기가 커짐

```
<html>
<body>
  <h2>웹 교육을 위한 서비스 기획안</h2>
  <h3>개요</h3>
  <p>
    웹 개발은 모바일이 중심이 된 IT시장에서도 여전히 중요한 위치에
있다.
  </p>
  <h3>개발</h3>
  <p>
    웹 개발은 주로 클라이언트와 백엔드 둘로 나눈다.
  </p>
  <h5>클라이언트 언어</h5>
  <ul>
    <li>HTML</li>
    <li>CSS</li>
    <li>JavaScript</li>
  </ul>
  <h5>백엔드 언어</h5>
  <ul>
    <li>PHP</li>
    <li>ASP</li>
    <li>Java</li>
    <li>Python</li>
  </ul>
</body>
</html>
```

웹 교육을 위한 서비스 기획안

개요

웹 개발은 모바일이 중심이 된 IT시장에서도 여전히 중요한 위치에 있다.

개발

웹 개발은 주로 클라이언트와 백엔드 둘로 나눈다.

클라이언트 언어

- HTML
- CSS
- JavaScript

백엔드 언어

- PHP
- ASP
- Java
- Python

HTML Tag - <a>…

하이퍼링크를 걸어주는 태그

```
<a href="http://www.naver.com">Go NAVER</a>
```

속성)

- href : 클릭시 이동 할 링크
- Target : 해당 링크를 보여줄 위치를 정하는 속성

_self : 현재 페이지 (기본값)
_blank : 새 탭
_parent : 부모 페이지로, iframe 등이 사용된 환경에서 쓰입니다.
_top : 최상위 페이지로, iframe 등이 사용된 환경에서 쓰입니다.
프레임이름 : 직접 프레임이름을 명시해서 사용할 수도 있습니다.

HTML Tag - <a>…

```
<html>
  <body>
    <a href="http://www.naver.com">Go Naver</a><br>
    <a href="http://google.co.kr" target="_blank">Go Google (new window)</a>
  </body>
</html>
```

Go Naver

Go Google (new window)

HTML Tag - <img>

이미지를 삽입하는 태그, src속성을 통해 이미지 경로를 지정

미지 파일이 src 속성에서 지정한 경로에 없을 시, 이미지는 출력되지 않거

나 엑스박스가 뜨게 됨.

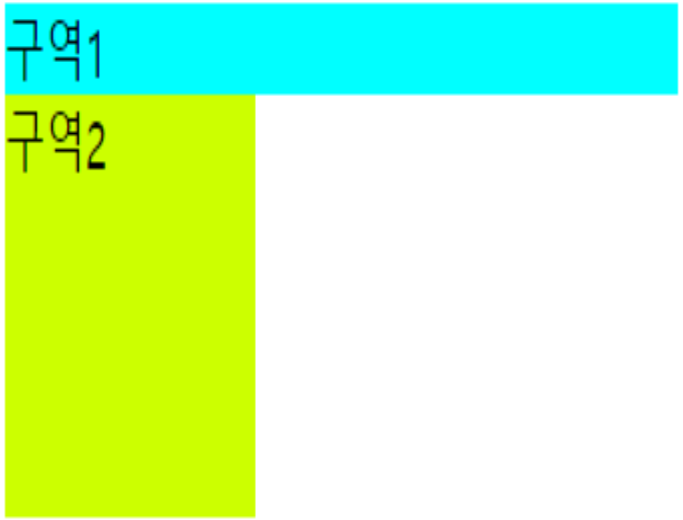
```
<html>
<body>
  <p>
    이미지가 정상적으로 삽입 된 경우<br>
    <img src="/images/attach/logo_black.png" width="245">
  </p>
  <p>
    없는 이미지가 삽입 된 경우<br>
    <img src="no_img.png" width="100" height="50">
  </p>
</body>
</html>
```

HTML Tag -

Division의 약자로, 레이아웃을 나누는데 주로 쓰임.

다른 태그와 다르게 특별한 기능을 가지지는 않음.

```
<html>
  <body>
    <div style="background-color:cyan">구역1</div>
    <div style="width:100px; height:100px; background-color:#CF0">구역2</div>
  </body>
</html>
```

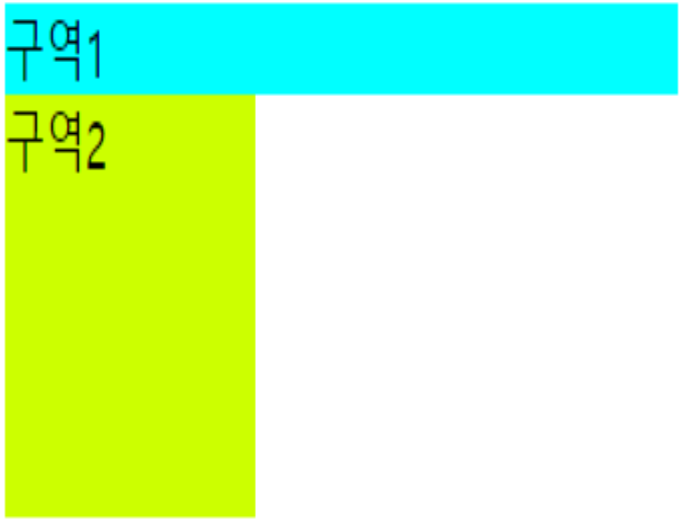


HTML Tag -

Division의 약자로, 레이아웃을 나누는데 주로 쓰임.

다른 태그와 다르게 특별한 기능을 가지지는 않음.

```
<html>
  <body>
    <div style="background-color:cyan">구역1</div>
    <div style="width:100px; height:100px; background-color:#CF0">구역2</div>
  </body>
</html>
```



HTML Tag -

 태그처럼 특별한 기능을 갖고 있지 않고,CSS와 함께 쓰임.

차이점은 div와 다르게 줄 바꿈이 되지 않음.

```
<html>
<body>
  <span style="background-color:red">span1</span>
  <span style="background-color:blue">span2</span>
  <span style="background-color:green">span3</span>
</body>
</html>
```

HTML Tag -

List의 약자로, 목록을 만다는 태그

 : 순서가 있는 (ordered list) 목록

 : 순서 없이 모양으로 (unordered list) 목록

```
<html>
  <body>
    <ol>
      <li>목록1</li>
      <li>목록2</li>
    </ol>

    <ul>
      <li>목록1</li>
      <li>목록2</li>
      <li>목록3</li>
      <ol>
        <li>목록3-1</li>
        <li>목록3-2</li>
      </ol>
    </ul>
  </body>
</html>
```

1. 목록1
2. 목록2

- 목록1
- 목록2
- 목록3
    1. 목록3-1
    2. 목록3-2

Thank you.

RNN 기초 / 류영표 강사
youngpyoryu@dongguk.edu