

Web Crawling

23.01.19 / 8기 유채원

CONTENTS

01. 데이터 수집

- Open API
- Crawling

02. 웹 기본 구조

- HTML/CSS/Javascript
- 개발자 도구
- 태그 경로 찾기

03. Beautiful Soup

- Beautiful Soup
- Find vs Select
- 알아야 할 함수

04. Selenium

- Selenium
- 알아야 할 함수

05. 활용 예시

06. 실습



EDA 일반론

데이터 수집

데이터 전처리

데이터
Scaling

데이터 시각화

사후 처리

오늘 세션 : <데이터 수집> 과정에 집중!

정형화된 데이터만 존재하지 않는다.

: 우리가 원하는 데이터는 항상 CSV 파일로만 존재하는 것이 아님 (ex. 국립 기상청 홈페이지의 일기 예보)

따라서 직접 정형화된 데이터를 만든다!

: 온라인에 퍼져 있는 여러 정보를 긁어모아 데이터화시키는 작업이 필수!

1. Open API

API (Application Programming Interface) : 애플리케이션이 요청과 응답을 주고받는 체계



Back-end

화면에 보여줄 정보를 처리
해주는 역할



Front-end

우리가 흔히 보는 화면
: 게시판의 틀, 제목 위치, 색상 및 글자 크기 등등

1. 데이터 수집

1. Open API



Back-end

1번 글에 대한 정보 요청



Database에서 1번 글에 대한

정보 찾아 제공



Front-end

사용자가 1번 게시물을 클릭!

Front-End가 정보를 요청할 때에 특정 규칙에 맞게 요청을 해야 함 = API (사용 규칙을 제공하는 것)

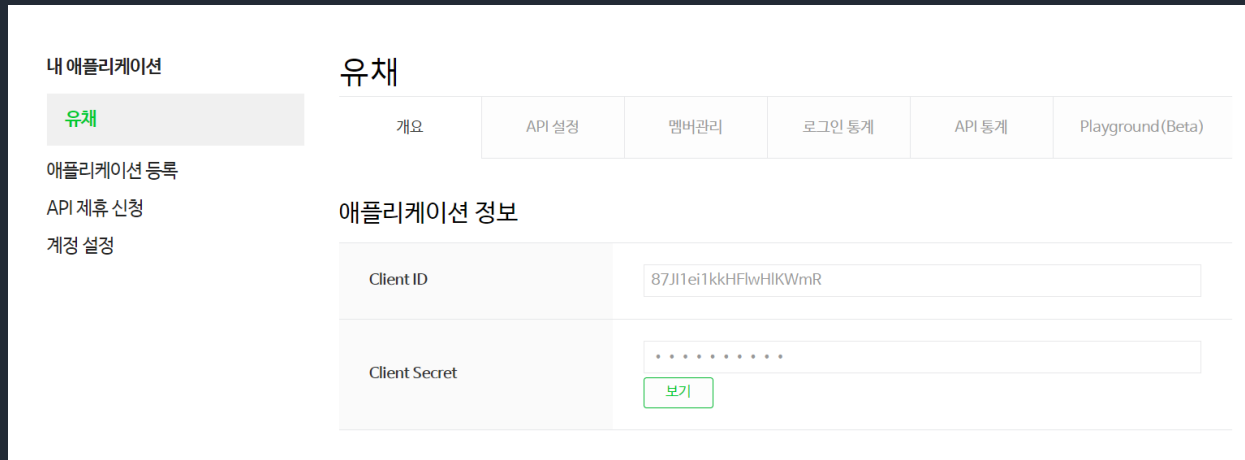
1. 데이터 수집

1. Open API

Open API: 누군가 백엔드를 만들어놓고 주소와 사용 규칙을 공개한 것 -> 누구나 사용 가능

Ex) 공공데이터포털(<https://www.data.go.kr/>), 네이버 오픈 API(<https://developers.naver.com/main/>)

1) 인증키 발급받기 (API 이용 신청)



The screenshot shows the Naver Open API application interface. On the left, under '내 애플리케이션' (My Application), the '유채' (Yuchae) tab is selected. Below it are links for '애플리케이션 등록' (Register Application), 'API 재휴 신청' (Apply for API Suspension), and '계정 설정' (Account Settings). The main area has a top navigation bar with '개요' (Overview), 'API 설정' (API Settings), '멤버관리' (Member Management), '로그인 통계' (Login Statistics), 'API 통계' (API Statistics), and 'Playground (Beta)'. The '앱리케이션 정보' (Application Information) section contains a 'Client ID' field with the value '87JI1ei1kkHFwHlKWmR' and a 'Client Secret' field with masked characters and a '보기' (View) button.

내 애플리케이션	
유채	
애플리케이션 등록	
API 재휴 신청	
계정 설정	


유채	
개요	API 설정
멤버관리	로그인 통계
API 통계	Playground (Beta)

앱리케이션 정보	
Client ID	87JI1ei1kkHFwHlKWmR
Client Secret 보기


1. 데이터 수집

1. Open API

2) 요청 URL(JSON 방식 추천),
요청 파라미터 (요청 시에 입력해야 할 값들) 확인하기

요청 URL 	
요청 URL	결괏값 반환 형식
<code>https://openapi.naver.com/v1/search/movie.xml</code>	XML
<code>https://openapi.naver.com/v1/search/movie.json</code>	JSON

요청 URL

파라미터 			
파라미터를 쿼리 스트링 형식으로 전달합니다.			
파라미터	타입	필수 여부	설명
query	String (필수)	Y	검색어. UTF-8로 인코딩되어야 합니다.
display	Integer	N	한 번에 표시할 검색 결과 개수(기본값: 10, 최댓값: 100)
start	Integer	N	검색 시작 위치(기본값: 1, 최댓값: 1000)
genre	String	N	영화 장르 코드 - 1: 드라마 - 2: 판타지 - 3: 서부 - 4: 공포 - 5: 로맨스 - 6: 모험 - 7: 스릴러 - 8: 느와르 - 9: 컬트 - 10: 다큐멘터리 - 11: 코미디 - 12: 가족 - 13: 미스터리 - 14: 전쟁 - 15: 애니메이션 - 16: 범죄

요청 파라미터

1. 데이터 수집

1. Open API

3) 정보 요청하기

http 요청 쉽게 보내주게 하는 패키지

```
import requests
client_Id = '87JI1ei1kkHF1wH1KWmR'
client_Secret = 'nKfSgRtopR'

naver_open_api = 'https://openapi.naver.com/v1/search/shop.json?query=android'
header_params= {'X-Naver-Client-Id':client_Id, "X-Naver-Client-Secret":client_Secret}
res = requests.get(naver_open_api, headers= header_params)

data = res.json()
```

(요청 URL)

(요청 파라미터)
파라미터 여러 개면 '&'로 연결

Json 파일 불러오기

1. 데이터 수집

2. Crawling

Crawling : 검색 엔진 로봇을 이용한 데이터 수집 방법

Web 상에 존재하는 Content들을 수집할 수 있음

- 1) 파이썬의 크롤러로 웹 서버에 정보 요청
- 2) 서버 응답을 받은 후 웹 서버와 상호작용하며 정보 획득
- 3) 얻은 정보를 핸들링하여 데이터화



웹의 기본 구조에 대해 알아야 함!

2. 웹 개발 언어 및 구조

웹 개발 언어에는?

HTML : 웹 브라우저에서 문서 및 웹 페이지가 표시되는 방법을 규정하는 언어

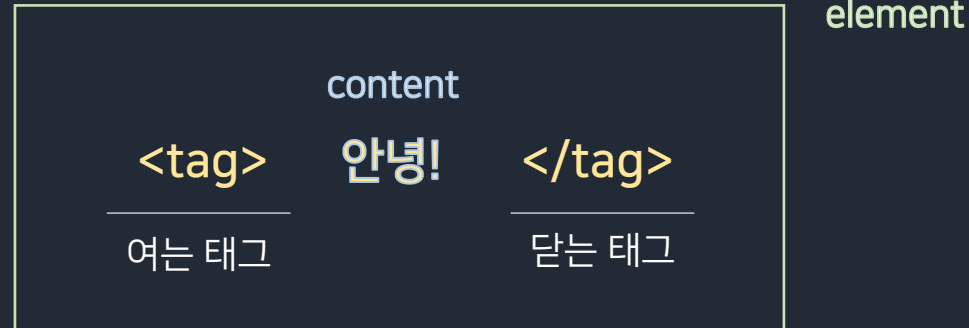
CSS : HTML로 만들어진 문서의 스타일을 지정하는 방식을 규정하는 스타일 시트 언어

Javascript : 웹 사이트에서 HTML과 CSS의 구성요소들의 동작을 변경할 수 있게 해주는 언어



2. 웹 개발 언어 및 구조

HTML - 웹 기본 구조



- 한 element는 여는 태그 + 내용 + 닫는 태그로 이루어져 있다.
- Tag끼리 상하 관계가 존재한다.

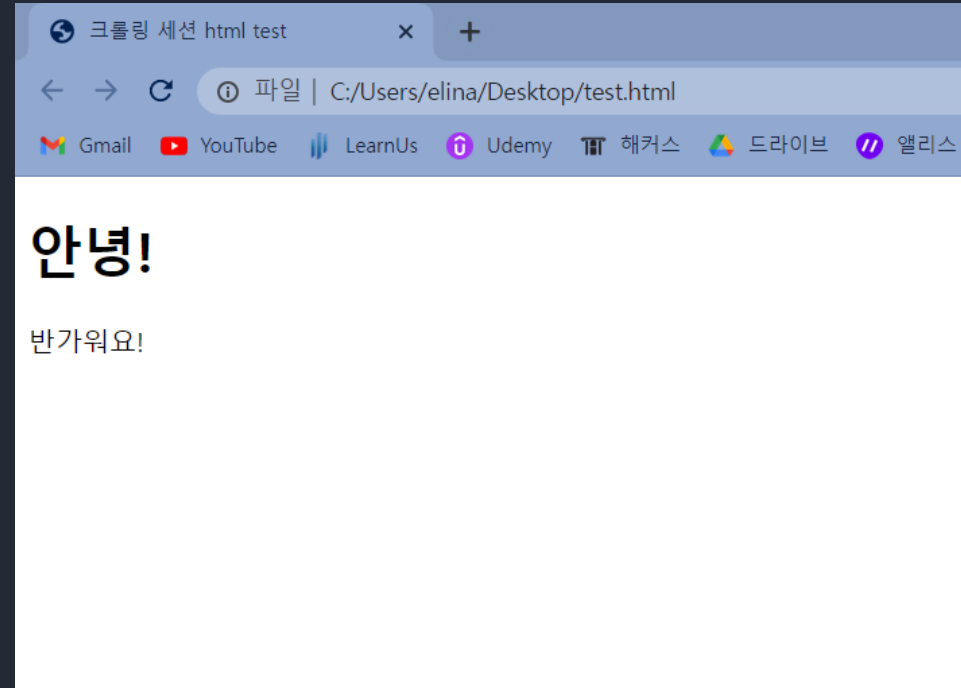
```
<h1>  
  <p> 안녕 </p>  
</h1>
```

: `<h1>` 태그 안에 `<p>` 태그가 있는 것!

2. 웹 개발 언어 및 구조

HTML - 웹 기본 구조

```
1 <!DOCTYPE html>
2 <html>
3   <head> <head> : 문서 전체에 대한 정보 (ex.제목)
4     <meta charset="utf-8">
5     <title>
6       크롤링
7     </title>
8   </head>
9   <body> <body> : 문서 실제 내용 시작
10    <h1> 안녕! </h1>
11    <p> 반가워요! 크롤링 세션..아직 한참 남았어요..ㅎㅎ</p>
12  </body>
13 </html>
```



- 우리가 흔히 크롤링하는 내용들은 주로 <body> 태그 안에 존재
- 태그 사이에 부모,자식,형제 관계가 존재하므로 내가 찾고자 하는 태그의 경로를 파악하는 것이 중요!

2. 웹 개발 언어 및 구조

CSS Selector

태그 속성

`<tag class = "abc" id = "content">` **안녕!** `</tag>`

동일한 tag들을
구분지어주는 역할

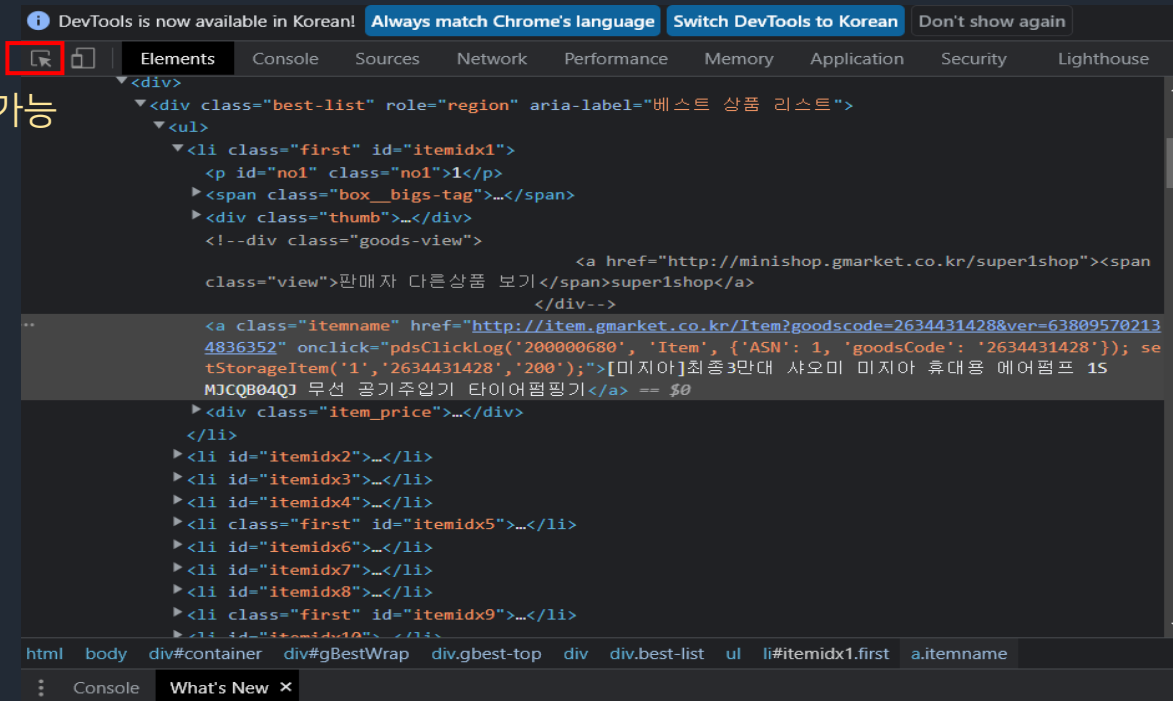
CSS Selector 이용하여 태그 경로 찾기

- 바로 하위의 태그(>): tag1 > tag2
- 하위의 태그(띄어쓰기) : tag1 tag2
- 클래스 이용하여 찾기(.) : tag1.abc
- Id 이용하여 찾기(.) : tag1#content

2. 웹 개발 언어 및 구조

개발자 도구

Element찾기 가능



1. 정보 추출하고자 하는 element에 마우스 대고 우클릭 - [검사]
2. 화면 우측 상단에 Chrome 맞춤 설정 - [도구 더보기] - [개발자도구]
3. 키보드 F12

2. 웹 개발 언어 및 구조

태그 경로 찾기

```
<div class="best-list" role="region" aria-label="베스트 상품 리스트">
  <ul>
    <li class="first" id="itemidx1">
      <p id="no1" class="no1">1</p>
      <span class="box__big-tag">...</span>
      <div class="thumb">...</div>
      <!--div class="goods-view">
        <a href="http://minishop.gmarket.co.kr/super1shop"><span
          class="view">판매자 다른상품 보기</span>super1shop</a>
      </div-->
      <a class="itemname" href="http://item.gmarket.co.kr/Item?goodscode=2634431428&ver=63809570213
        4836352" onclick="pdsClickLog('200000680', 'Item', {'ASN': 1, 'goodsCode': '2634431428'}); se
        tStorageItem('1', '2634431428', '200');">[미지아]최종3만대 샤오미 미지아 휴대용 에어컨프 15
        MJCQB04QJ 무선 공기주입기 타이어펌핑기</a> == $0
      <div class="item_price">...</div>
```

1. 눈으로 구조 파악해서 경로 찾기 : 복수의 데이터 뽑을 때 유용 (`div.best-list div.thumb a`)
2. Element 마우스 우클릭 - [Copy] - [Copy Selector] : 단일 데이터 뽑을 때 유용

3. Beautiful Soup

Beautiful Soup

Request 패키지로 텍스트 형태의 html 문서 가져온 후,
이 텍스트 형태에서 원하는 html 태그를 추출할 수 있게 해줌 -> Beautiful Soup

```
import requests
from bs4 import BeautifulSoup

res = requests.get(웹페이지 주소)
Soup = BeautifulSoup(res.content, 'html.parser')
items = soup.find(태그 경로) 텍스트 형태의 html문서
```


3. BeautifulSoup

Find vs Select

Select -> CSS Selector로 태그 객체를 찾아 반환

단일추출 : `find = select_one` : 가장 처음 찾은 태그 객체 반환

Ex) `<div class = 'abc'>` 하위 태그인 `` 찾기

- `soup.find('div', attrs = {'class' : 'abc'}).find('a', attrs = {'id' : 'best'})`
- `Soup.select_one('div.abc a#123')`

Find의 경우 반복적으로 코드를 작성해야 함.

다중추출 : `find_all = select` : 해당하는 모든 태그들을 찾아 list 형태로 반환

```
items = soup.select('div.abc')
```

```
for item in items:
```

```
    print(item.get_text)
```

3. BeautifulSoup

알아야 할 함수

웹사이트 제목 가져오기: `soup.find('title')`

`item = soup.select_one(태그 경로)`

- `item.attrs`
 - `{'class': ['itemname'],`
`'href': 'http://item.gmarket.co.kr/Item?goodscode=2634431428&ver=638095766565048545',`
`'onclick': "pdsClickLog('200000680', 'Item', {'ASN': 1, 'goodsCode': '2634431428'}); setStorageItem('1','2634431428','200');"}]`
- `Item['class']`
- `Item['href']` : 태그 내의 하위 링크 주소
- `Item.get_text()` : content (태그 안의 내용) 가져오기
 - '[미지아]샤오미 미지아 휴대용 에어펌프 1S MJCQB04QJ 무선 공기주입기 타이어펌핑기'

Selenium : 동적 웹 스크래핑 방법

- Javascript가 동적으로 만든 데이터를 크롤링할 수 있게 해줌
 - 사용자가 실제로 브라우저를 탐색하는 것처럼 작동
 - HTML 요소 클릭, 페이지 이동, 키보드 입력 등이 자율적
 - (장점) 사용자에게 제공되는 눈에 보이는 콘텐츠는 전부 크롤링 가능!
 - (단점) But 브라우저를 직접 켜서 움직이므로 시간이 오래 걸린다.
- > requests 패키지와 같이 사용해 속도를 줄이는 방법

Chromedriver 설치

Chromedriver 설치 : 작성된 코드를 webdriver를 통해 브라우저로 전달하는 역할

크롬 버전 확인 : chrome://version 주소창에 입력

크롬드라이버 다운로드 링크 : chromedriver.chromium.org/downloads

개인의 버전에 맨 앞 숫자와 동일한 드라이버 다운로드 받기!

```
from selenium.webdriver.common.by import By
```

```
단일추출 : driver.find_element(By.CSS_SELECTOR, "태그 경로")
```

```
다중추출 : driver.find_elements(By.CSS_SELECTOR, "태그 경로")
```

여러 가지 함수

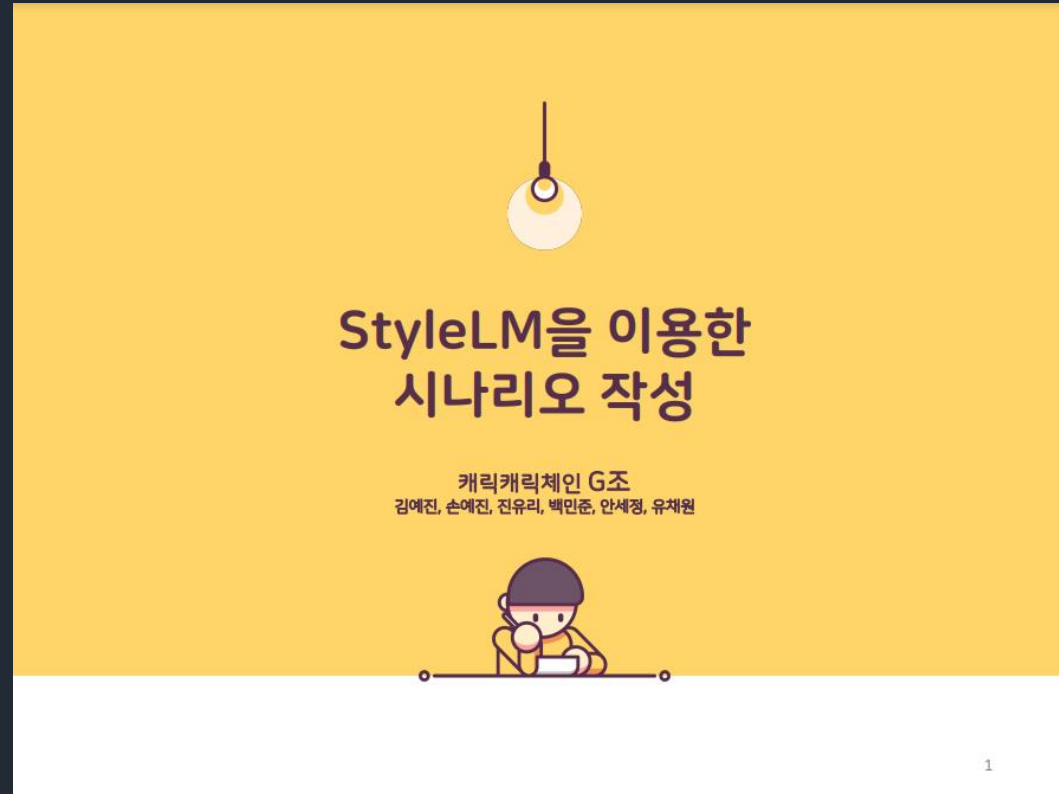
브라우저 탭 이동

- `driver.back()` : 뒤로 가기
- `driver.forward()` : 앞으로 가기
- `driver.quit()` : 종료
- `driver.execute_script('window.scrollTo(0, document.body.scrollHeight);')` : 스크롤 내리기

태그 관련

- `tag.click()` : 클릭
- `tag.send_keys('검색어')` : 검색어 입력
- `tag.send_keys(Keys.ENTER)` : 엔터

5. 활용 예시






시나리오 대사 필요!

5. 활용 예시

미스터션샤인 시나리오 받을 수 없을까요?	 까치산
파수꾼  14	 감라봉
변산	 감라봉
어린 의뢰인 오리지널 각본	 초록우
영화 <교회누나> 시나리오	 kissgood
영화"대관람차" 시나리오	 xeva
더킹	 감라봉
지구를 지켜라	 감라봉



변산 Sunset in My Hometown, 2017
관람객      **8.34** | 기자·평론가      **6.50** | 네티즌      **7.72** | 내 평점
드라마 | 한국 | 123분 | 2018.07.04 개봉 | [국내] 15세 관람가
감독 이준익 출연 박정민(학수), 김고은(선미) 더보기 >

다운로드  2,206    

주요정보 배우/제작진 포토 동영상 평점 리뷰 명대사/연관영화

배우



박정민
주연 | 학수 역
밀수, 2021
언프러입드, 2021



김고은
Kim Go-eun
주연 | 선미 역
인택트, 2020
영웅, 2020

Selenium:

각 게시물에 차례대로 들어가 파일 다운 받게 하기

여러 페이지 크롤링 : for문 이용

게시물 태그. click() 이용

Beautiful Soup:

크롤링한 영화의 장르, 주요 배역명 뽑기

6. 실습

첨부 파일을 확인해주세요!

DATA

SCIENCE LAB

발표자 유채원 010-8736-1815
E-mail: elinafe71@gmail.com