K-Digital Training 스마트 팩토리 3기

데이터 크롤링

데이터 크롤링이란?

웹 사이트에서 정보를 수집하는 것

	A	B	C
1	상품명	판매가	url
2	(내맘대로 10봉씩섞어 30봉) 김치라면 진라면 신라면	19,900	http://item.gmarket.co.kr/Item?goodscode=635868496
3	오뚜기 진라면 매운맛 120gX40봉 한박스	40,200	http://item.gmarket.co.kr/Item?goodscode=1865938970&buyboxtype=ad
4	컵라면 30개세트 라면 10종x3개씩 농심 오뚜기 팔도	44,900	http://item.gmarket.co.kr/Item?goodscode=2845309383&buyboxtype=ad
5	오뚜기 진라면 매운맛 120gX5봉	5,200	http://item.gmarket.co.kr/Item?goodscode=1865940943&buyboxtype=ad
6	오뚜기 진라면 순한맛 120gX5봉	5,200	http://item.gmarket.co.kr/Item?goodscode=1865945369&buyboxtype=ad
7	오뚜기 진라면 매운맛 120gx40봉지 / 한박스	29,800	http://item.gmarket.co.kr/Item?goodscode=2648808801&buyboxtype=ad
8	오뚜기 진라면 순한맛 (봉지/5입)	5,700	http://item.gmarket.co.kr/Item?goodscode=2925023164&buyboxtype=ad
9	오뚜기 진라면 매운맛 (큰컵/박스/12입)	19,000	http://item.gmarket.co.kr/Item?goodscode=2923049582&buyboxtype=ad
10	농심 신라면 (봉지/5입) / 봉지라면모음전 여러가지 20여종	7,000	http://item.gmarket.co.kr/Item?goodscode=2925036649&buyboxtype=ad
11	엔즈몰/컵라면 모음/진라면/너구리/신라면/사발면	35,000	http://item.gmarket.co.kr/Item?goodscode=2242471146&buyboxtype=ad
12	오뚜기)진라면매운맛작은컵(15개입)	27,840	http://item.gmarket.co.kr/Item?goodscode=1954346532&buyboxtype=ad
13	오뚜기 진라면 순한맛 (소컵/박스/6입) / 농심 오뚜기 삼양 팔도 컵라면 모음	7,600	http://item.gmarket.co.kr/Item?goodscode=2920670928&buyboxtype=ad
14	오뚜기 진라면 순한맛 (봉지/5입) / 브랜드 봉지라면 20여종 모음	5,700	http://item.gmarket.co.kr/Item?goodscode=2925018163&buyboxtype=ad
15	진라면 신라면 짜파 해물라면 스낵 삼양 x30봉(각5봉)	25,900	http://item.gmarket.co.kr/Item?goodscode=2018125832
16	오뚜기 진라면 순한맛 120gx40봉지 / 한박스	29,800	http://item.gmarket.co.kr/Item?goodscode=2650346501&buyboxtype=ad
17	오뚜기 진라면 매운맛 (소컵/박스/6입)	7,600	http://item.gmarket.co.kr/Item?goodscode=2920671156&buyboxtype=ad
18	오뚜기 진라면 매운맛 (봉지/5입)	5,700	http://item.gmarket.co.kr/Item?goodscode=2925028313&buyboxtype=ad
19	진라면 순한맛 40봉입 오뚜기 봉지라면	31,900	http://item.gmarket.co.kr/Item?goodscode=1656635368&buyboxtype=ad

HTML

HTML(Hyper-Text Markup Language)

- 웹 사이트를 만들 때 사용되는 마크업 언어

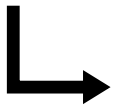
HTML 태그

- 웹 페이지를 구성하는 기본적인 구성 요소

– HTML 문서에서 요소의 성격을 지정하거나 페이지의 구조를 결정하는 역할을 한다.

HTML 태그

네이버로 이동



a : 태그 명
href : 속성 (태그의 종류에 따라 사용할 수 있는 속성이 다르다)

태그의 종류

- 제목 태그 : h1 ~ h6
- 본문 태그 : p
- 문단 태그 : div
- 링크 태그 : a
- 목록 태그 : ul, ol, li
- 이미지 태그 : img
- 입력 태그 : input

```
<div>                    <!-- h1과 ul의 부모 태그 -->
  <h1>제목<h1/>          <!-- div의 자식 태그 -->
  <ul>                   <!-- div의 자식 태그이자 li의 부모 태그 -->
    <li>목록1</li>       <!-- ul의 자식 태그 -->
    <li>목록1</li>
    <li>목록1</li>
  </ul>
</div>
```

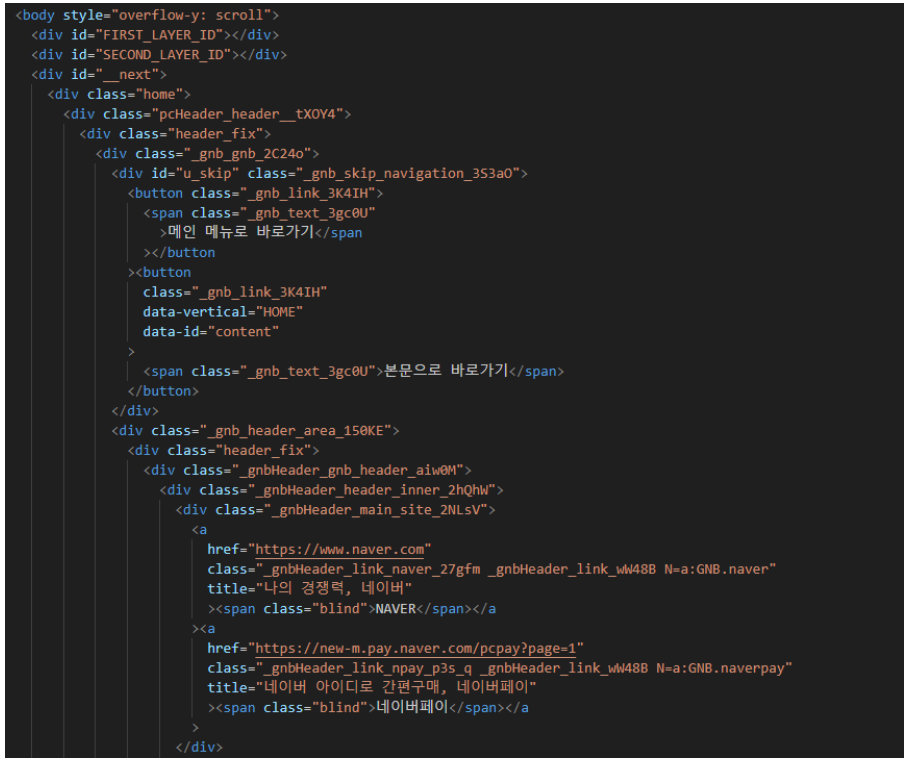
requests

requests

- HTTP 프로토콜을 이용하여 웹 사이트로부터 데이터를 송수신 할 수 있는 Python 라이브러리

```
import requests
```

```
response = requests.get("https://shopping.naver.com/home")
html = response.text
print(html)
```



beautifulsoup

beautifulsoup

- 웹 페이지의 HTML, XML 파일에서 데이터를 추출하는 Python 라이브러리
- HTML 태그를 검사하고 선택할 수 있다.

```
!pip install bs4

from bs4 import BeautifulSoup
```

```
soup = BeautifulSoup(html, 'html.parser')
```

CSS 선택자

CSS 선택자

CSS(Cascading Style Sheets) 선택자

- HTML 요소를 스타일링하거나 특정 요소를 선택할 때 사용한다.

클래스 선택자

- ‘class’ 속성을 기반으로 요소를 선택한다.
- ‘.클래스명’

아이디 선택자

- ‘id’ 속성을 기반으로 요소를 선택한다.
- ‘#아이디명’

openpyxl

openpyxl

- 파이썬에서 엑셀 파일을 쉽게 다루기 위한 오픈 소스 라이브러리
- 엑셀 파일 형식을 읽고 쓰는 기능을 제공하며, 파이썬 코드를 사용하여 엑셀 파일의 데이터를 조작할 수 있다.

```
!pip install openpyxl
```

openpyxl

```
import openpyxl

# 엑셀 만들기
wb = openpyxl.Workbook()

# 워크시트 만들기
ws = wb.create_sheet('codingon')

# 데이터 추가하기
ws['A1'] = '이름'
ws['B1'] = '영어이름'

ws['A2'] = '김세령'
ws['B2'] = 'sarah'

# 엑셀 저장하기
wb.save('/content/drive/MyDrive/python_lecture/codingon.xlsx')
```

```
# 파일 불러오기
wb = openpyxl.load_workbook(r'/content/drive/MyDrive/python_lecture/codingon.xlsx')

# 시트 선택
ws = wb['codingon']

# 파일 수정하기
ws['A3'] = '김소연'
ws['B3'] = 'lily'

# 파일 저장하기
wb.save('/content/drive/MyDrive/python_lecture/codingon.xlsx')
```