
심화 교육과정
데이터 수집, 전처리 – 크롤링, 전처리기법
THINK LIFE SYNC AI
Infinyx

I 웹 크롤링

파이썬 크롤링 하는 이유는 무엇인가 ?

딥러닝 학습을 위해서는 이미지 데이터 혹은 수치화 데이터가 필요한 경우가 있습니다.

API 호출이 가능한 데이터가 있는 반면 API 공개하지 않은 경우도 있습니다.

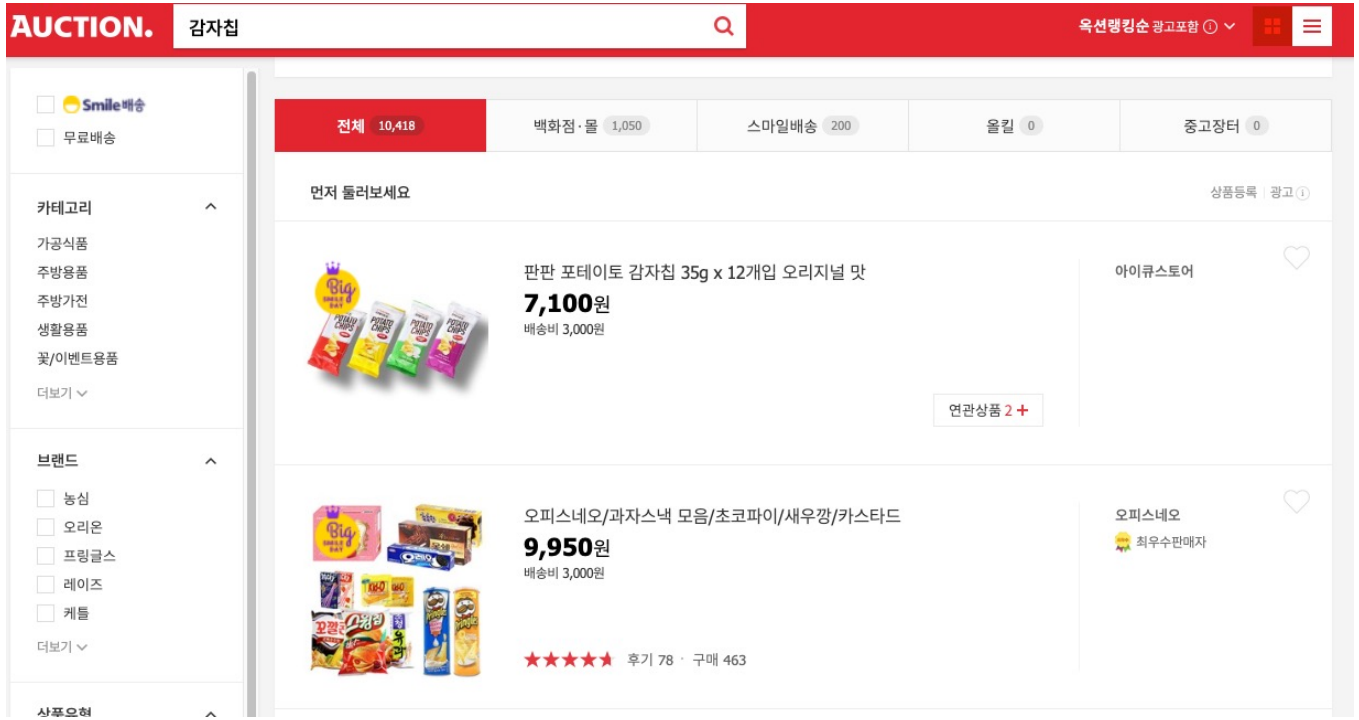
즉 크롤링은 자신의 원하는 정보를 웹페이지에서 가져오는 행위 또는 작업의 의미로 일컬어 집니다.

크롤링 예시

1. abc.com
2. bcd.or.kr
3. abcde.com
4. test.com

흔한 예시로 구글 검색 엔진에 우리가 검색해서 나오는 정보들도 모두 구글의 봇이 크롤링 하여서 얻어진 정보입니다.

크롤링 예시



상업적으로 경쟁사의 제품 정보나 A 마켓의 제품 정보를 크롤링 하여 사용하는 것

참고로 상업적으로 사용하려면 정보를 제공하는 측에 허가 필요합니다.

크롤링 예시

XML JSON 광주광역시_광산구_태양광발전설치현황

활용신청 오류신고 및 담당자 문의

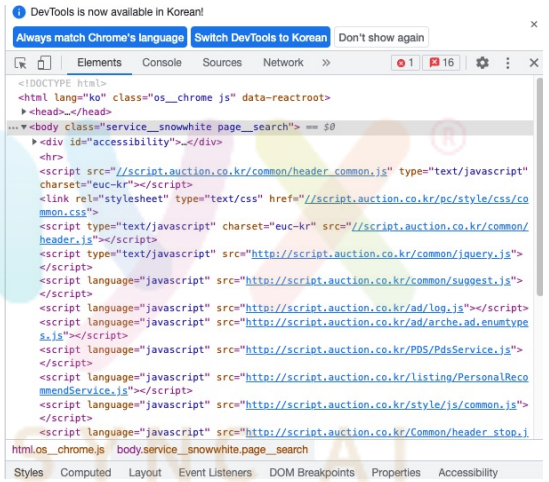
오픈API 정보 메타데이터 다운로드

서비스	광주광역시_광산구_태양광발전설치현황_20211231		
분류체계	환경 - 환경일반	제공기관	광주광역시 광산구
관리기관	공공데이터활용지원센터	관리기관 전화번호	1566-0025
보유근거		수집방법	
업데이트 주기	연간	차기 등록 예정일	2023-05-16
매체유형	텍스트	전체 행	2189
확장자	XML, JSON	활용신청	0
데이터 한계		키워드	태양광발전소,신재생에너지,태양열
등록	2022-05-16	수정	2022-05-16

합법적인 정보 제공원(API 등)에 정보 전달의 한계나 유용하지 않아서 추가적으로 정보 획득이 필요한 경우

Infinyx

웹 크롤링의 프로세스 설명



크롤링 도구 가져오기

웹 사이트
구글, 네이버 등

1. 필요한 정보가 있는 페이지 구역을 선택합니다.

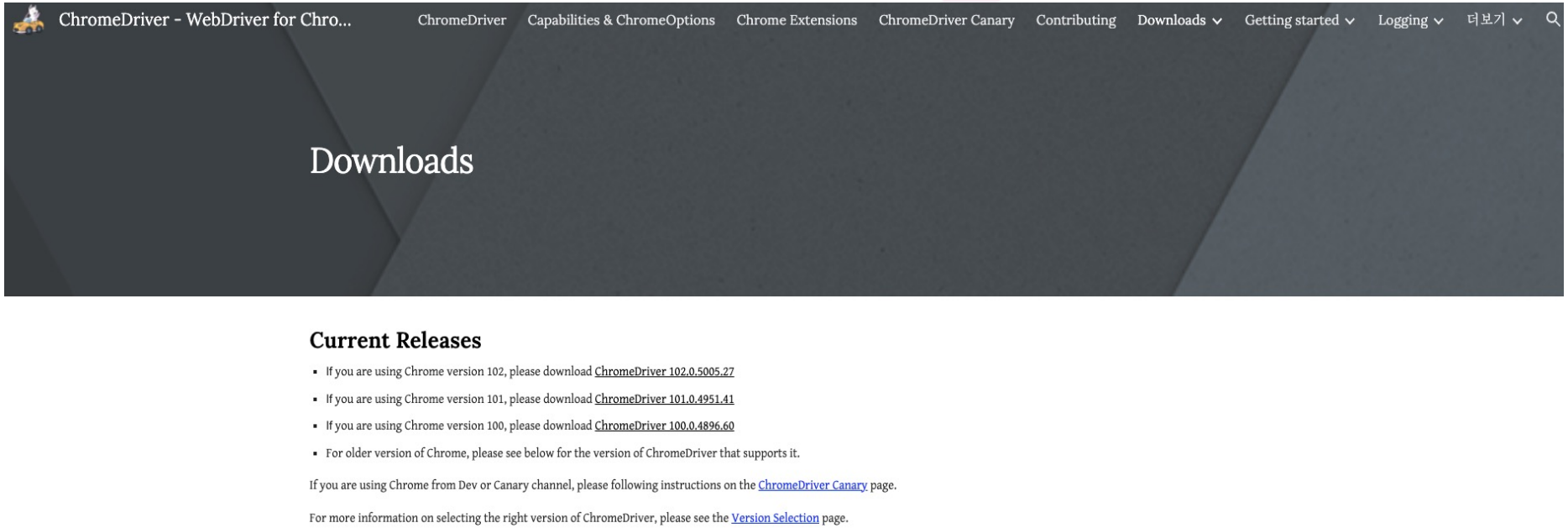
2. 선택한 구역이나 페이지의 데이터 가져옵니다.

데이터 저장

크롤링의 흐름

- 크롤링 하는 도구는 아주 종류가 다양합니다. (requests, scrapy, selenium 등)
- 본 강의에서 사용할 도구는 파이썬 selenium을 이용합니다.

- 필요 드라이브 다운로드 https://chromedriver.chromium.org/downloads
- 자신 PC 설치된 Chrome 버전 확인 필수



용어 정리

- 셀레늄(selenium) 본래는 원격으로 웹브라우저를 컨트롤하려는 목표로 만들어 졌으나 크롤링에도 쓰이는 도구
- HTML : 웹페이지 상의 데이터들을 글자형식의 코드로 표현한 형태의 언어
- HTML 태그 : 위의 HTML 코드 중에서 여러 성격의 데이터들을 책갈피로 찾을 수 있게 표시한 값
- 엘리먼트 : HTML 태그를 이용해 찾은 데이터

크롤링의 실습

- 실습 프로세스

1. 크롤링 도구 코드에 띄우기
2. 오픈 마켓접속 -> 예제 –> 옥션
3. 상품검색 -> 음식(아무거나)
4. 상품리스트 가져오기
5. 상품리스트에서 필요한, 상품명, 가격, 상품링크 출력

크롤링의 실습

- 실습 프로세스

1. 크롤링 도구 코드에 띄우기
2. 구글 검색 창
3. 검색 -> 동물명
4. 동물 이미지 링크 가져오기

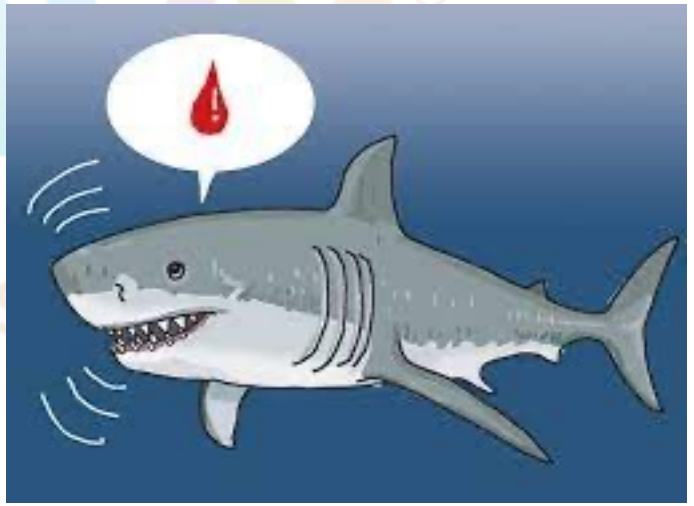
데이터 정재 작업

- 이미지 전처리 : 이미지라는 데이터는 단편적으로 보이지만, 실제로 매우 많은 데이터를 가짐 따라서 이를 한 번 가공해주는 과정을 거쳐야만 컴퓨터가 좀 더 효율적으로 분석할 수 있다.

- 즉 이미지내 에서 불필요한 데이터를 줄이고 유의미한 데이터를 정제하는 과정이 필요 이러한 과정을 전처리 알고리즘이라고 부릅니다.

간단히 말하면 이미지를 조작하는 과정, 이를 통해 딥러닝 모델의 정확도 즉 성능을 높일 수 있다.

데이터 정재 작업 – 크롤링, 수집 데이터 불필요한 데이터 제거 작업



- 실제 Real world 에 속한 이미지로 학습 필요
- 이미지 화질, 사람이 눈으로 봐도 구분이 안되는 경우 제거 필수
- 오른쪽 사진처럼 애니메이션 혹은 일러스트 합성 이미지 제거 필요

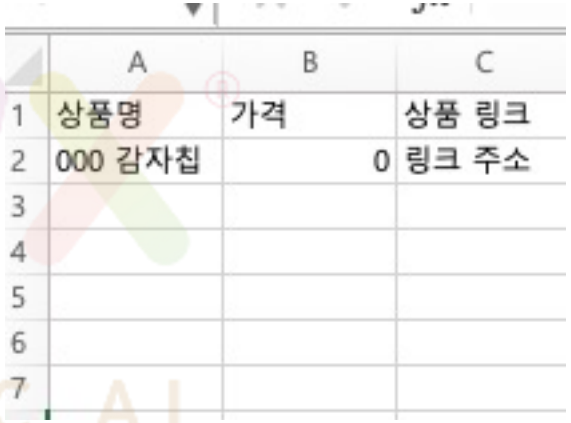
데이터 정재 작업 – 이미지 사이즈 변경, 회전, 반전, 색상 변환

- 필요한 이미지 사이즈로 변경, 혹은 회전, 아편, 반전, 색상 변환 등 다양한 기법을 활용하여 이미지 정재

크롤링의 과제

- 과제 프로세스

1. 크롤링 도구 코드에 띄우기
2. 오픈 마켓접속
3. 상품검색 -> 음식(아무거나)
4. 상품리스트 가져오기
5. 상품리스트에서 필요한, 상품명, 가격, 상품링크 출력
6. 상품명, 가격, 상품 링크 -> .CSV 저장하기



	A	B	C
1	상품명	가격	상품 링크
2	000 감자칩	0	링크 주소
3			
4			
5			
6			
7			

- CSV 양식 예시

컬럼명 상품명, 가격, 상품 링크

제출 방법 : 코드 제출

크롤링의 실습

- 실습 프로세스

1. 크롤링 도구 코드에 띄우기

2. 구글 검색 창

3. 검색 -> 동물명

4. 동물 이미지 링크 가져오기

5. 동물 이미지를 각 동물 명칭 폴더에 저장

동물 제시 : 상어 , 돌고래 , 고래


감사합니다.
THINK LIFE SYNC AI
Infinyx