

파이썬을 이용한 데이터수집 및 스마트공장 견학

Crawling 활용 2

2021년 1월 17일
안재관

금일 목표

- 일반적인 crawling 절차를 숙지한다
- 아이템의 selector 검색 시 검색 범위를 박스로 좁히는 중요성 파악
- 예외 처리의 필요성 파악
- 파일 읽기/쓰기 방법 숙지
- 다나와 크롤링 예제를 통한 전처리 배우기

일반적인 Crawling 절차

- 1. HTML 소스 가져오기 (selenium or requests)
- 2. 추출할 아이템(들)을 정하고, 이를 공통적으로 포함하는 박스 정하기
- 3. 박스의 selector 파악하기
- 4. 아이템(들)의 selector 파악하기
- 5. for문 내에서 각 박스별로 아이템 추출 및 출력/저장
- 6. 필요시 다음 페이지/키워드로 넘어가기

● 예시 : <https://blog.naver.com>

● 전체 코드 (한 페이지)

```
from selenium import webdriver
import time
browser = webdriver.Chrome("./chromedriver")

browser.get("https://blog.naver.com")
time.sleep(1)
boxes = browser.find_elements_by_css_selector(".info_post")
for box in boxes:
    item1 = box.find_element_by_css_selector("em.name_author").text
    item2 = box.find_element_by_css_selector("a strong").text
    item3 = box.find_element_by_css_selector(".like em").text
    print(item3, "\t", item1, "\t", item2)
```











일반적인 Crawling 절차

- 1. HTML 소스 가져오기 (selenium or requests)
- 2. 추출할 아이템(들)을 정하고, 이를 공통적으로 포함하는 박스 정하기
- 3. 박스의 selector 파악하기
- 4. 아이템(들)의 selector 파악하기
- 5. for문 내에서 각 박스별로 아이템 추출하기
- 6. 필요시 다음 페이지/키워드로 이동하기

● 예시 : <https://blog.naver.com>

● 결과

15	뿌요	충청남도농업기술원 향지촌 아이
18	안다송이	배도라지즙 비타민디 자연농
74	공취생	[부산 기장] 오리백숙이 맛있
29	허니	이동갈비 마당소 소갈비 집에서
123	쿠니	강원도 여행 마치고 청평역
66	담덕공자	하와이안 스타일 광고 맛집
14	샤넬에스터 바비	세포라 투페이스드
30	차니	엄마선물 추천 엘클라 타임리스
16	삐약이	가평 가족펜션 샤갈의 마을
25	별이	특별한 카네이션선물 이수역꽃집

 뿌요 10시간 전 충청남도농업기술원 향지촌 아이 갑자기 날씨가 좀 더워졌습니다.5 여름 시원한 음료를 찾는데향지촌 디.250ml향지촌은 ... 공감 15 댓글 8	 허니 10시간 전 이동갈비 마당소 소갈비 집 매 끼니 어떤 음식을 차려야 할까늘 비를 집에서 구워먹었어오고기반찬 런 구워... 공감 29 댓글 7	 샤넬에스터 바비 11시간 전 세포라 투페이스드 행오버 안녕하세요 ~~브랑브랑왁싱브로: 세팅스프레이입니다!!픽서로 유문 에 투페이스드 행오버 ... 공감 14 댓글 3	 별이 11시간 전 특별한 카네이션선물 안녕하세요!! 별이가 왔떠음 무 바빠서 어버이날도 열령 카네이션선물을 해드리려고 공감 25 댓글 7
 안다송이 10시간 전 배도라지즙 비타민디 자연농 기관지가 약한 우리 가족은 환절기 아 괜찮네요.미세먼지도 작년보다 습니다.거기다 햇빛 썬기 힘들... 공감 18 댓글 3	 쿠니 10시간 전 강원도 여행 마치고 청평역 강원도 여행을 마치고 저녁 식사 하 니 여기 드림 닭갈비가 짜잔 ~ 간판 부근 닭갈비 맛집이라고 하... 공감 123 댓글 6	 차니 11시간 전 엄마선물 추천 엘클라 타임리스 ELLCLA TIMELESS TREASURER 구매한 엘클라 트레저 리주비네이 핑몰에서 구매... 공감 30 댓글 5	
 공취생 10시간 전 [부산 기장] 오리백숙이 맛 오랜만에 드라이브도 할겸 몸보신 집인만수장가든을 소개한다.만수: 가게의 크기가 엄청나게 넓다.입... 공감 74 댓글 26	 담덕공자 10시간 전 하와이안 스타일 광고 맛집 광고 맛집 봉주르 하와이 내부 수리: 렸네요. 핫한 홍대 1호점에 이어 2호 을 했습니다...^^이곳에 ... 공감 66 댓글 6	 삐약이 11시간 전 가평 가족펜션 샤갈의 마을 안녕하세요 삐약이입니다삐약악 : 아직은 코로나가끝나지 않아서 늘 은 연휴를마무리하며 가족과... 공감 16 댓글 7	


일반적인 Crawling 절차

● 1. HTML 소스 가져오기 (selenium or requests)

- selenium : 오래 걸리지만 브라우저에서 검사 기능을 사용하기 좋음
- requests : 빠르지만 동적 웹페이지 처리 불가
- 어떤거 쓸지 판단? : 페이지 소스 보기에 내가 원하는 콘텐츠가 나오면 requests

● 2. 추출할 **아이템**(들)을 정하고, 이를 공통적으로 포함하는 **박스** 정하기


- 박스 : 추출할 아이템들을 모두 담고 있는 가상의 네모
- 4단계에서 아이템 추출을 용이하게 해줌

1  **아이빙빙**
10시간 전

2 **닥터지 블랙 스네일 크림 이랑 같이 잘쓰는 잇템!!**

악건성 + 민감성 피부를 가지고 계신 분들 기초 어떤 것 사용하고 계시나요? 저는 진짜 안써본 게 없을 정도로 거의 다 써본것 같아요. 백화점 고가 브랜드는 물론이고 로드샵까지 다른 분들이 좋다고 했던 것은 거의 다 써보는 편인데 제...

3 공감 37 댓글 13

1  **더핏카샵**
8시간 전

2 **아우디 A5!! 내 생애 첫차를 더핏카샵으로 보내뵙니다**

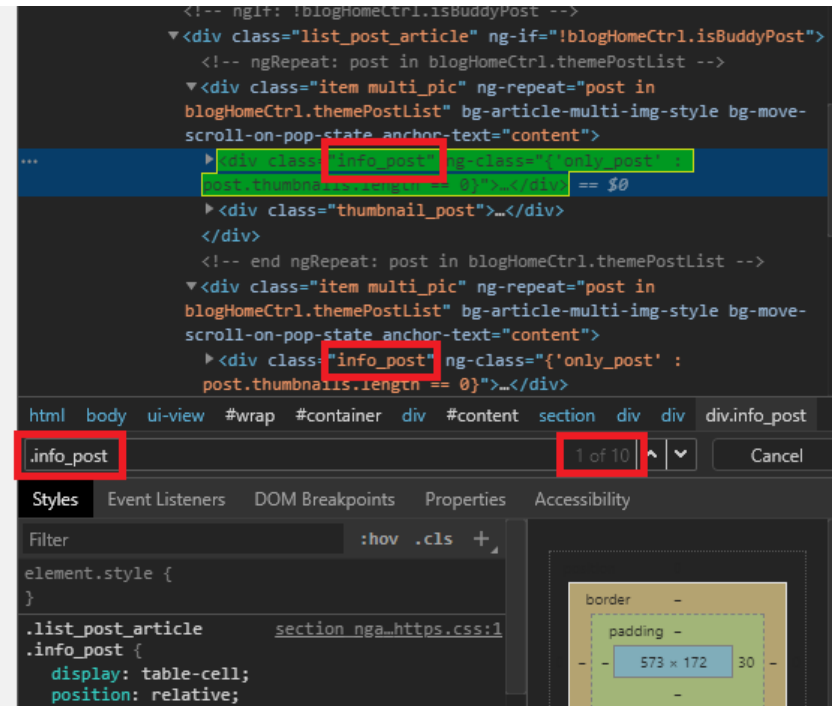
짠!~안녕하세요!! 더핏카샵 마스터짠~하고 나타나!! 인사드립니다벌써 1월 중순^^내 이럴줄 알았지~지 난간 세월 되돌리수는 없기에후회없는 하루하루가 되시길오늘도 마스터 최선을 다해 보겠습니다여러분들도 봐이팅!!2020...

3 공감 20 댓글 14

일반적인 Crawling 절차

● 3. 박스의 selector 파악하기

- 웹페이지에서 **눈으로 박스가 몇 개** 있는지 파악
- **검사에서 selector로 검색**했을 때 검색 결과와 비교
- .info_post로 검색하면 10개 나옴
- 포스트 개수도 10개



일반적인 Crawling 절차

● 4. 아이템(들)의 selector 파악하기

- 박스 전후의 소스코드는 무시하고 파악하기
- 즉, 박스가 잘 잡혔다는 가정하에 검사에서 selector로 검색했을 때 여러 개가 나와도 웬만하면 상관없음
- 포스트 제목 selector : a strong 으로 추출 가능
- 검사에서 a strong으로 검색하면 27개 나옴

a strong 4 of 27

```
<div class="info_post" ng-class="{ 'only_post' : post.thumbnails.length == 0 }">
  <a ng-href="https://blog.naver.com/rejio" class="author" target="_blank" bg-nclick="out*1.profile"
    href="https://blog.naver.com/rejio">
    <div class="thumbnail_author">...</div>
    <div class="info_author">...</div>
    ::after
  </a>
  <!-- ngIf: blogHomeCtrl.loggedIn && !post.buddyRelationType && !post.isBuddyWithUser -->
  <div class="desc">
    <a ng-href="https://blog.naver.com/rejio/221772577273" class="desc_inner" target="_blank" bg-nclickf="{ 'code': 'out*1.text', 'cid': '90000003_0000000000000033A2ACF9F9', 'rank': 1 }" href="https://blog.naver.com/rejio/221772577273">
      <strong class="title_post" ng-bind-html="post.noTagTitle || post.title">닥터지 블랙 스네일 크림 미
      랑 같이 살쓰는 잇템!!</strong>
    </a>
    <a ng-href="https://blog.naver.com/rejio/221772577273" class="text" ng-bind-html="
      'post.briefContents || post.contents' target="_blank" bg-nclickf="{ 'code': 'out*1.text', 'cid':
      '90000003_0000000000000033A2ACF9F9', 'rank': 1 }" ng-show="post.briefContents || post.contents" href=
      'https://blog.naver.com/rejio/221772577273">...</a>
    </div>
  <div class="comments">...</div>
</div>
```

- 하지만 상관없음, 왜? 박스 내에서는 유일한 a strong 이기 때문

일반적인 Crawling 절차

● 4. 아이템(들)의 selector 파악하기

```

7 boxes = browser.find_elements_by_css_selector(".info_post")
8 print(len(browser.find_elements_by_css_selector("a strong")))
9 for box in boxes:
10     print(len(box.find_elements_by_css_selector("a strong")))

```

- 7라인 : .info_post 로 찾은 박스들을 boxes에 저장
- 8라인 : **웹페이지 전체에서 a strong** 을 만족하는 태그가 몇 개냐
- 9라인 : 모든 박스들에 대해 반복해라
- 10라인 : **각 박스 내에서 a strong** 을 만족하는 태그가 몇 개냐
- 결과 : 우측
 - 웹페이지 전체에는 a strong이 27개
 - 검사로 검색한 결과와 동일
 - 박스 10개 모두 a strong이 1개씩 있음
- 박스로 분리할 경우 박스 내 아이템 selector 찾기가 훨씬 쉬워짐
- 각 아이템을 웹페이지 전체가 아니라 박스 내에서만 골라내면 되기 때문
- 코드가 간단해지고 추후 유지/보수 쉬워짐



27
1
1
1
1
1
1
1
1
1
1
1

일반적인 Crawling 절차

- 5. for문 내에서 각 박스별로 아이템 추출 및 출력/저장
- 6. 필요시 다음 페이지/키워드로 넘어가기

예외 및 오류 처리

● try-except-else)

- Crawling시 오류가 언제 발생할지 예측 불가능
- 예외 처리를 하지 않을 경우에는 프로그램이 중단됨
- 예외 처리를 통해 예외가 있더라도 프로그램이 끝까지 실행될 수 있도록 함
- 사용법

try:

일단 이 부분을 시도해봄

except:

try문에서 에러 발생시 이 부분 실행

else:

try문에서 에러 미발생시 이 부분 실행 (생략 가능)

- 예시 : 산기대 일반 계약학과에서 영문 학과명 없는 경우
 - <http://www.kpu.ac.kr/contents/main/cor/sanhak.html>

예외 및 오류 처리

● try-except-else

- 예시 : 산기대 일반 계약학과에서 영문 학과명 없는 경우
- <http://www.kpu.ac.kr/contents/main/cor/sanhak.html>

```

browser.get("http://www.kpu.ac.kr/contents/main/cor/sanhak.html")
time.sleep(1)
boxes = browser.find_elements_by_css_selector(".unit")
numSuccess = 0
numFailure = 0
for box in boxes:
    try:
        item1 = box.find_element_by_css_selector(".ko div").text
        item2 = box.find_element_by_css_selector(".en div").text
        item3 = box.find_element_by_css_selector(".homepage").get_attribute("href")
    except:
        print("skipping this box")
        numFailure += 1
    else:
        print(item1, "\t", item2, "\t", item3)
        numSuccess += 1
print("성공 :", numSuccess, "실패 :", numFailure)
browser.close()

```

```

기계제조공학과      MANUFACTURING ENGINEERING      http://subweb.
기계설계·시스템공학과      PRODUCTION AND MECHANICAL DESIGN E
메카트로닉스시스템공학과      MECHATRONICS SYSTEM ENGINEERING
컴퓨터융합공학과      COMPUTER CONVERGENCE      http://subweb.
부품소재공학과      MATERIALS AND COMPONENTS ENGINEERING ht
환경안전경영학과      DEPARTMENT OF ENVIRONMENT AND SAFETY M
화합물반도체공학과      MAJOR IN LIGHT EMITTED DIODE ENGINEERI
기업경영학과      COOPERATE MANAGEMENT      http://subweb.kpu.
스마트컴퓨터융합공학과      SMART COMPUTER CONVERGENCE DEPARTM
skipping this box
성공 : 9 실패 : 1

```

기계제조공학과 [HOMEPAGE](#)
MANUFACTURING ENGINEERING

기계설계·시스템공학과 [HOMEPAGE](#)
PRODUCTION AND MECHANICAL DESIGN
ENGINEERING

메카트로닉스시스템공학과 [HOMEPAGE](#)
MECHATRONICS SYSTEM ENGINEERING

컴퓨터융합공학과 [HOMEPAGE](#)
COMPUTER CONVERGENCE

부품소재공학과 [HOMEPAGE](#)
MATERIALS AND COMPONENTS
ENGINEERING

환경안전경영학과 [HOMEPAGE](#)
DEPARTMENT OF ENVIRONMENT AND
SAFETY MANAGEMENT

화합물반도체공학과 [HOMEPAGE](#)
MAJOR IN LIGHT EMITTED DIODE
ENGINEERING

기업경영학과 [HOMEPAGE](#)
COOPERATE MANAGEMENT

스마트컴퓨터융합공학과 [HOMEPAGE](#)
SMART COMPUTER CONVERGENCE
DEPARTMENT

계약학과 대학원 [HOMEPAGE](#)

파일 읽기/쓰기

● 파일 읽기

- 검색어 목록 등이 텍스트 파일에 저장되어 있을 경우....등
- 예시

```
f = open("keywords.txt", "r", encoding="utf8")
for line in f:
    print(line, end="")
f.close()
```

포켓몬
산기대
도시락

Process finished with exit code 0

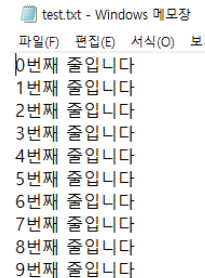
- r : 파일을 읽기 (read) 용도로 열기
- encoding : 한글이 깨질 경우가 있음
- 텍스트 파일을 리스트로 간주하면 요소 하나하나는 텍스트 한 줄이 됨
- print 함수에 end 옵션을 주면 줄 끝에 이를 출력함, end 옵션의 기본값에는 "wn" 가 들어가서 자동으로 줄바꿈이 되는 효과인데, end=""는 줄바꿈을 하지 말라는 뜻임, line의 끝에 이미 "wn"이 있기 때문
- 경로를 적어야 할 경우 \ 대신 / 또는 \\ 사용
 - O : f = open("C:/Users/cnc4e/PycharmProjects/kpu/keywords.txt", "r", encoding="utf8")
 - O : f = open("C:\\Users\\cnc4e\\PycharmProjects\\kpu\\keywords.txt", "r", encoding="utf8")
 - X : f = open("C:\Users\cnc4e\PycharmProjects\kpu\keywords.txt", "r", encoding="utf8")
- 파일 사용 완료 후에는 꼭 .close()

파일 읽기/쓰기

● 파일 쓰기

- Crawling 결과를 화면에 계속 출력하다보면 메모리 부족 현상 발생 가능
- 화면에만 결과 출력시 프로그램이 실수로 종료되면 결과물 분실됨
- 프로그램을 오래 돌려야 하는 경우에는 결과를 필히 파일에 기록
- 예시

```
f = open("test.txt", "w")
for i in range(10):
    f.write(str(i) + "번째 줄입니다\n")
f.close()
```



- w : 파일을 쓰기 (write) 용도로 새로 열기, **기존에 파일이 있었으면 지워짐, 하루 내내 프로그램 돌려서 파일에 쓰게 했다가, 파일을 백업하지 않고 그대로 새로 돌리면 기존 파일 없어지니 꼭 유의해야함**
- a : 파일의 뒤쪽에 이어쓰기 (append) 용도로 새로 열기
- .write() : print() 함수처럼 쓰면 됨, 대신 따옴표 밖의 쉼표는 인자 구분으로 인식하기 때문에 쓰면 안됨, 마지막에 줄바꿈을 위해 \n 필수

날짜 처리

● 필요성

- ISO 8601 국제 날짜 표준은 YYYY-MM-DD임, 예) 2020-01-17
- Crawling 하다보면 온갖 날짜 표시 형식을 마주침, 시간까지 필요하면....
 - 네이버뉴스 2020.01.17. 오전 9:41
 - 다음뉴스 2020.01.17. 10:55
 - 동아일보 2020-01-17 03:00
 - 세계일보 2020-01-15 18:37:06

● datetime

- 현재시각, 날짜출력, 시간출력

```
from datetime import datetime
x = datetime.now()
print(x)
print(x.date())
print(x.time())
```

```
2020-01-17 11:19:06.366544
2020-01-17
11:19:06.366544
```

날짜 처리

● 기본 날짜설정

- 시간/분/초 설정하지 않으면 기본값 0으로

```
x = datetime(2020, 5, 20)
print(x)
```

2020-05-20 00:00:00

● 형식에 따른 출력 방법

```
print(x.strftime("%d.%m.%Y"))
print(x.strftime("%Y-%m-%d"))
```

20.05.2020

2020-05-20

● 문자열을 형식에 따라 읽기

```
str = "09/19/2018"
date = datetime.strptime(str, "%m/%d/%Y")
print(date.date())
```

2018-09-19

● 날짜 대소비교

```
print(datetime.now() < date)
print(datetime.now() > date)
```

False

True

날짜 처리

● 시간변경

```
from datetime import timedelta
date = datetime(2020, 2, 1)
for i in range(5):
    date += timedelta(days=1)
    print(date.date())
```

```
2020-02-02
2020-02-03
2020-02-04
2020-02-05
2020-02-06
```


Pandas/openpyxl 패키지

- 구글에서 "파이썬 엑셀 파일" 검색시 openpyxl, pandas 패키지가 많이 등장함
- 읽기는 pandas 패키지 이용
- 추출하기는 openpyxl 패키지 이용(anaconda 사용시 설치 생략 가능)
- 파일 읽고 쓰기

```
import pandas as pd
```

```
dt = pd.read_excel('./titanic.xlsx')  
Print(dt)
```

- 파일 내보내기

```
# ! Pip install openpyxl  
Dt.to_excel('./titanic_2.xlsx', index=False)
```

Pandas/openpyxl 패키지

- 파일 내보내기

```
Dt.to_excel('./titanic_2.xlsx', index=False)
```

실습

● 산기대 – 학사공지

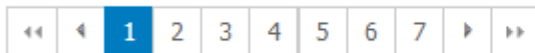
- 주소 : <http://www.kpu.ac.kr/contents/main/cor/noticehaksa.html>
- 대상 : 검색어 "입학" 으로 나오는 글 중 공지가 아닌 70개의 게시글
- 수집내용 : 번호, 제목, 작성자, 등록일, 조회수
- 저장 : kpu_haksa_입학.xlsx로 저장

• 팁

- 페이지를 넘겨도 주소가 변하지 않음

kpu.ac.kr/contents/main/cor/noticehaksa.html

kpu.ac.kr/contents/main/cor/noticehaksa.html



- 페이지 번호를 우클릭 – 새 탭에서 열기
- 숨어있던 제대로된 주소가 나옴

kpu.ac.kr/front/boardlist.do?currentPage=3&menuGubun=1&siteGubun=14&bbsConfigFK=1&searchField=D.TITLE&searchValue=%C0%D4%C7%D0&searchLowItem=ALL#

- 주소에 페이지번호와 검색어를 직접 넣을 수 있음

실습: 다나와 크롤링

● 다나와 무선 청소기 데이터 수집

- 과정 (pseudocode)
 - 주소 얻기
 - '무선청소기' 검색하기(방법은 두가지)
 - 무선 청소기 "박스", "이름", "스펙", "가격" selector 찾기
 - 함수 만들어보기
 - for문을 사용하여:
 - 수집내용들을 각각 변수에 저장하기
 - 수집 데이터가 빈 경우 건너뛰기
 - 엑셀 파일에 쓰기

The screenshot shows the Danawa website interface. At the top, there's a search bar with the text '무선청소기' (Wireless Vacuum Cleaner) and a search icon. Below the search bar, there are navigation links like '카테고리' (Category), '통합검색' (Integrated Search), '상품' (Product), '여행' (Travel), '장터' (Market), '쇼핑가이드' (Shopping Guide), '뉴스' (News), '커뮤니티' (Community), '리뷰' (Review), '동영상' (Video), and '다나와Q&A' (Danawa Q&A). The main content area displays a list of vacuum cleaners with their images, names, and prices. For example, one product is '삼성전자 Gen11 20V 무선 필리핀 청소기' (Samsung Gen11 20V Wireless Philippines Vacuum Cleaner) priced at 89,000원. Another is '다이슨 무선 진공 청소기 8TB-VC120 (홍갈색)' (Dyson Wireless Vacuum Cleaner 8TB-VC120 (Red/Grey)) priced at 198,000원. The sidebar on the left contains filters for '카테고리' (Category), '제조사별' (By Manufacturer), '브랜드별' (By Brand), '품목' (Item), '모터' (Motor), '출입력' (Input Power), and '사용시간(계급)' (Usage Time (Class)). The bottom of the page shows a summary of the search results, including the total number of items found (321,640) and a list of filters applied.

최근

광고상행 >

특정 키워드나 품목/해외직구 제외,
품질 상품 제외, 중복품 검색을 이용하세요.
다시보기 >

40%

데이터 수집 후 전처리 하기

● 다나와 무선 청소기 데이터 불러오기

- Pandas로 엑셀 파일 불러오기
 - Info() 함수
 - Head(), tail() 함수

```

1 # 예제 다나와 크롤링 결과 가져오기
2 import pandas as pd
3 data = pd.read_excel('./danawa_crawling_result_2022_1.xlsx')
4 data.info()
5 data.head(20)

```

	상품명	스펙 목록	가격
0	LG전자 오브제컬렉션 코드제로 ThinQ A9S AO9571	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입+걸레겸용 / 소비전력: 590W...	1055760
1	삼성전자 비스포크 제트 VS20A956A3	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입+걸레겸용 / [성능] 흡입력: ...	587830
2	LG전자 코드제로 ThinQ A9S AS9370IKT	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입+걸레겸용 / 소비전력: 590W...	814610
3	샤오미 CLEANFLY 차량용 무선 청소기 4세대 H2 (해외구매)	차량용청소기 / 무선 / 흡입력: 16,800Pa / 최대출력: 120W / 헤파필...	60590
4	샤오미 TROUVER POWER 11	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입전용 / 소비전력: 400W / ...	73950
5	B1vVXWAHbT	B1vVXWAHbT483019	483019
6	NaN	NaN	0
7	베이스어스 차량용 청소기 A3 (해외구매)	차량용청소기 / 무선 / 흡입력: 15,000Pa / 최대출력: 135W / 헤파필...	50350
8	샤오미 드림미 V10	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입전용 / 소비전력: 450W / ...	141435
9	삼성전자 비스포크 제트 VS20A957E3	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입+걸레겸용 / [성능] 흡입력: ...	812900

데이터 수집 후 전처리 하기

● 유용한 pandas 전처리 함수

- dropna(): column 내에 NaN 값이 있으면 해당 내용은 필요없다 간주하고 삭제
- Reset_index(): 다중 인덱스 데이터 프레임의 인덱스를 재설정

```

1 # 예제 다나와 크롤링 결과 가져오기
2 import pandas as pd
3 data = pd.read_excel('./danawa_crawling_result_2022_1.xlsx')
4 data.info()
5 data.head(20)

```

	상품명	스펙 목록	가격
0	LG전자 오브제컬렉션 코드제로 ThinQ A9S AO9571	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입+걸레검용 / 소비전력: 590W...	1055760
1	삼성전자 비스포크 제트 VS20A956A3	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입+걸레검용 / [성능] 흡입력: ...	587830
2	LG전자 코드제로 ThinQ A9S AS9370IKT	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입+걸레검용 / 소비전력: 590W...	814610
3	샤오미 CLEANFLY 차량용 무선 청소기 4세대 H2 (해외구매)	차량용청소기 / 무선 / 흡입력: 16,800Pa / 최대출력: 120W / 헤파필...	60590
4	샤오미 TROUVER POWER 11	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입전용 / 소비전력: 400W / ...	73950
5	B1vVXWAHbT	B1vVXWAHbT483019	483019
6	NaN	NaN	0
7	베이스어스 차량용 청소기 A3 (해외구매)	차량용청소기 / 무선 / 흡입력: 15,000Pa / 최대출력: 135W / 헤파필...	50350
8	샤오미 드리미 V10	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입전용 / 소비전력: 450W / ...	141435
9	삼성전자 비스포크 제트 VS20A957E3	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입+걸레검용 / [성능] 흡입력: ...	812900

데이터 수집 후 전처리 하기

● 유용한 pandas 전처리 함수

- If 문과 drop() 사용하여 잘못 크롤링 된 열 제거.

예시:

```
idx=[]
for i in range(len(data_2['상품명'])):
    if len(data_2['상품명'][i].split()) <= 1:
        idx.append(i)

print(len)
```

1	data_2.head(50)				
38	41	샤오미 드림미 V11SE	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입전용 / 소비전력: 450W / ...	179900	
39	42	삼성전자 비스포크 제트 VS20A957D3P	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입+걸레겸용 / [성능] 흡입력: ...	810790	
40	43	오토모 ATM-H100	핸디/스틱청소기 / 핸디형 / 무선형 / 흡입전용 / 소비전력: 70W / [배터리] ...	31330	
43	46	샤오미 SHUNZAO 차량용 무선청소기 2세대 Z1 (해외구매)	차량용청소기 / 무선 / 흡입력: 7,000Pa / 최대흡력: 90W / 2종필터 ...	25130	
44	47	m8q6EZFGXq	m8q6EZFGXq708939	708939	
45	48	일렉트로룩스 WELL Q7 WQ71-2ESSF	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입전용 / [성능] 싸이클론 / [...	193930	
46	49	삼성전자 제트 VS15R8543Q4CW	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입전용 / 소비전력: 410W / ...	464070	
47	51	한샘 트리플 플러스 2.0 QNBC-6000W	육실청소기 / 핸디+스틱형 / 무선형 / [배터리] 충전시간: 4시간 / 사용시간(...	65940	
48	52	샤오미 TROUVER POWER 12	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입전용 / 소비전력: 450W / ...	170190	
49	53	LG전자 코드제로 A9 A9100S	핸디/스틱청소기 / 핸디+스틱형 / 무선형 / 흡입+걸레겸용 / 소비전력: 450W ...	604360	

```
1 idx=[]
2 for i in range(len(data_2['상품명'])):
3     if len(data_2['상품명'][i].split()) <= 1:
4         idx.append(i)
5
1 len(idx)
4
1 idx
[5, 44, 83, 125]
1 data_3=data_2.drop(index=idx, inplace=False)
2
```