

크롤링_프로젝트_7

프로젝트 결과물

유지은_2021120119
김하늘_2022111733
이승윤_2023111759
김유경_2023111780
박정원_2023111791

I 크롤링 과정

1. 크롤링 조건을 맞추기 위한 방법 및 코드 설명

이번 실습 프로젝트에서 주어진 크롤링 조건은 1) '아웃터' 항목에서 '자켓' 분류로 카테고리를 설정하기 2) '리뷰 많은순'으로 분류된 상품들을 필터링하기 3) 필터링된 상위 30개의 자켓 상품 이었습니다. 위 조건들에 해당하는 웹페이지의 URL을 지정했습니다.

```
# Chrome 웹 브라우저를 실행하기 위해 webdriver.Chrome()을 사용합니다.
# chromedriver 실행 파일 경로를 인자로 전달합니다.
browser = webdriver.Chrome()
# 브라우저를 이용하여 지정한 URL을 엽니다.
browser.get(
    "https://zigzag.kr/categories/-1?title=%EC%9D%98%EB%A5%98&category_id=-1
    &middle_category_id=436&sub_category_id=438&sort=201")
time.sleep(1)

# 현재 웹 페이지의 소스 코드를 가져와서 변수 'html'에 저장합니다.
html = browser.page_source
# BeautifulSoup을 사용하여 HTML 소스 코드를 파싱합니다.
soup = BeautifulSoup(html, "html.parser")
```

한편, 3)의 조건을 맞추기 위해 `soup.select('.e1dr6ufx0 > a')[:30]`을 이용하여 html 문서에서 클래스가 'e1dr6ufx0'인 요소의 하위에 있는 요소들 중에 30개 제품을 선택했습니다.

```
# 상위 30개 아웃터 제품 정보 구간 선택
zigzag_post_area = soup.select('.e1dr6ufx0 > a')[:30]
```

2. 크롤링 요소를 수집하기 위한 방법 및 코드 설명

수집해야 하는 첫 번째 크롤링 요소는 상품 정보입니다. 이때 상품의 제목, 상품의 가격과 할인율, 리뷰 평점 및 개수, 상품의 썸네일이 해당 상품 정보에 포함됩니다. 우선, 크롤링 요소를 구하기 전에 다음과 같이 목표 결과값을 넣을 리스트를 초기화했습니다.

```
result = []
review_result = []
```

상품의 제목, 가격, 리뷰 평점 및 개수를 구할 때에는 반복문을 사용하여 각 포스트에 대한 정보를 추출했습니다. BeautifulSoup의 `select_one` 메서드를 사용하여 html 문서에서 특정 클래스에 해당하는 요소를 선택했습니다. 이후 `.text`를 이용하여 선택된 요소를 텍스트 형태로 추출했습니다.

상품의 할인율을 구하는 과정에서, 상품별로 할인이 적용되는 요소와 할인이 적용되지 않은 요소가 있었습니다. 후자의 경우 할인율로 정한 변수 `discount_rate`에 해당하는 값이 없기 때문에 `None`값으로 처리하였습니다. 이를 `if`문을 사용해 구분하였습니다.

```
# 상위 30개 아우터 제품에 대한 데이터 추출
for zigzag_post in zigzag_post_area:
    title = zigzag_post.select_one('.e91dh064').text # 제목
    price = zigzag_post.select_one('.eh5ooyt0').text # 가격

    discount_rate = zigzag_post.select_one('.e91dh062') # 할인율
    # 할인하는 제품과 하지 않는 제품으로 나누어져 있어 조건문 활용
    if discount_rate is not None:
        discount_rate = discount_rate.text
    else:
        discount_rate = "None"

    rating = zigzag_post.select_one('.e91dh069 .e13zfay41').text # 평점
    reviews_number_element = zigzag_post.select_one('.e13zfay40') # 리뷰
    # 개수의 element를 가져옴
    reviews_number_text = reviews_number_element.text
    reviews_number = int(''.join(filter(str.isdigit,
    reviews_number_text))) # 리뷰 개수 - 문자로 되어있어서 정수로 변환
```

상품의 썸네일은 이미지 형식이기때문에 선택된 이미지 태그에서 `'src'` 속성을 통해 썸네일 이미지의 URL을 가져옵니다. 여기서 추출된 이미지 URL을 변수 `thumbnail`에 저장했습니다.

```
thumbnail_element = zigzag_post.select_one('.e81k49g1 > .e81k49g0 > div
> img') # 썸네일 이미지 URL
thumbnail = thumbnail_element.get('src')
```

이후 추출된 요소를 `append()` 함수를 사용하여 만들어둔 리스트에 추가합니다.

```
result.append([title, price, discount_rate, rating, reviews_number, thumbnail])
```

수집해야 하는 두 번째 크롤링 요소는 상품 리뷰입니다. 상품 상세페이지 내 리뷰탭로 이동하여 전체 리뷰 중 상위 5개 리뷰의 1)리뷰어이름, 2)리뷰날짜, 3)리뷰텍스트를 구해야 합니다.

우선, 하위페이지인 각 상품에 해당하는 리뷰 페이지로 이동하는 코드를 작성하여 리뷰 정보를 수집하였습니다.

```
for product_link in zigzag_post_area:
    product_url = "https://zigzag.kr" + product_link['href']
    browser.get(product_url)
    time.sleep(1)
```

리뷰들 중 일정 글자수를 초과하여 '더보기'버튼을 클릭해야 리뷰 원본이 나오는 경우가 있습니다. 이를 예외 처리를 통해 해결하기 위해 다음과 같이 try에 실행할 코드를 넣고 except에는 예외가 발생했을 때 처리하는 코드를 넣었습니다.

```
# 더보기 버튼 클릭 -> 리뷰 펼치기
try:
    while True:
        more_button = browser.find_element(By.CSS_SELECTOR,
            '.BODY_13.BOLD.css-1aa4nqt.eox2jl01') # 더보기 버튼 찾기
        browser.execute_script("arguments[0].click();", more_button) # 해당
        버튼을 찾으면 JavaScript를 실행시켜 클릭
        time.sleep(2)
# 더보기 버튼 없는 리뷰
except:
    pass # 루프 종료하고 리뷰 그대로 출력

product_html = browser.page_source
product_soup = BeautifulSoup(product_html, "html.parser")
```

다음으로, 각 상품 페이지에서 반복문을 사용하여 상위 5개 리뷰 데이터를 추출했습니다.

```
# 각 상품 페이지에서 상위 5개 리뷰 데이터 추출
review_info_area = product_soup.select('.e1hi9732')[:5]

for review_info in review_info_area:
    reviewer_name = review_info.select_one('.e1fnwskn0').text # 리뷰어 이름
    review_date = review_info.select_one('.e1okf4zi0').text # 리뷰 날짜
    review_text = review_info.select_one('.eox2jl02').text.replace('\n', ' ')
# 리뷰 텍스트 - 엔터키를 사용해 입력한 데이터로 인해 개행문자 사용하여 출력
```

추출된 리뷰 요소들을 append() 함수를 사용하여 만들어둔 리스트에 추가합니다.

```
review_result.append([reviewer_name, review_date, review_text])
```

마지막으로, 결과값을 데이터 프레임으로 변환 후 결과값을 csv 파일로 저장합니다.

```
print(result)
print(review_result)

# 결과값을 데이터프레임으로 변환
```


	reviewer_name	review_date	review_text
1	0.**	22.03.14	고고싱 옷은 늘 가격대비 퀄리티가 좋다고 생각합니다. 자켓도 생각보다 퀄리티있었습니다 다만 옷 상태가 그래도 돈주고 산건데. 먼지 투성이에 새상품 아닌 중고장에서 주워온것같은 상태로
2	le**	23.11.03	키가 적다보니 큰 자켓은 싫어서 고민을 많이 했어요. 이 제품은 제가 본 자켓중에서도 작은 편인데 가격까지 너무 착해서 믿어도 되나 싶었어요. 입어봤는데 전체적으로 작아서(특히 팔) 안에
3	jj**	23.09.13	허리라인이 있는지 몰랐는데 살짝 허리라인이 있어서 오버핏, 캐주얼보다 약간 청장같은 느낌이었어요. 그래서 좀 아쉽기는한데 그래도 가격대비 만족해요^^
4	br**	22.05.16	요즘 자켓이 사고싶어져서 여기저기 찾아봤는데 , 제구가 적다보니 입으면 죄다 아빠양복 자켓 크기라 고민하고 있었는데 역시 고고싱 ㉡ 평소애 이 사이트가 저랑 맞는게 많아 자주 시키는데
5	es**	22.04.21	솔직히 재질이 좋단애기 못하겠지만 이만한 가격에 이런 자켓 구하기는 힘들듯.. 안에 두꺼운 옷 입어도 넉넉할것같아서 좋아요 세모로 떨어지는 끝부분 핏은 조금 아쉬운데 팔 안벌릴때 또
6	do**	23.10.08	기본 검은 블레이저가 활용도가 좋아서 사고싶어서 이 기회에 구매했습니다! 솔직한 후기는요, 키150 몸무게 40한테 큰 오버핏이어서 캐주얼하게 입기에는 길이가 괜찮고, 저는 조금 길이를 줄
7	bg**	23.10.06	와... 이 가격이 이 퀄리티 맞아요? 진짜 알을 것 같고 별 기대 안 했는데 생각보다 퀄리티가 너무 좋아요!!!! 실밥 빠져나온 것도 하나 없고 진짜 다 완벽해요 길이도 160 기준 엉덩이 바로 밑! 딱
8	yh**	21.09.26	재질도 엄청 탄탄하고 요즘같은 일교차 큰 날씨에 입기 딱 좋아요 키168인데 엉덩이 반정도 가려집니다! 소매는 팔이 긴분들은 좀 짧다고 느껴질실 거 같아요! 저는 딱 괜찮았습니다!! 입었을
9	da**	23.10.02	정장을 입는 타임은 아닌데 나이를 먹으면서 단정한 자켓이 꼭 필요해서서 구매 하게 됐어요 블랙 블레이저는 워낙 무난템이다보니 요거서 살지 저기서 살지 엄청 고민 했는데 사길 잘 했다고
10	m**	23.11.01	급하게 필요해서 산 자켓인데 키 156 기준 엉덩이 살짝 가려요 입었을때 단정하고 깔끔해 보이구요!! 손등은 살짝 가리는게 조금 커보이긴 하는데 오버핏 다른 자켓들보단 제 체구에 찰 나아요
11	da**	23.08.08	딱 원하던 색깔이에요! 너무 밝은 회색은 원하지 않았는데 짙은 검정에 가까운 차콜이라 좋았습니다! 기장도 좋고 품도 여유롭게 잘 맞아요. 포장상태도 좋았는데 처음에 냄새가 너무 역해서 좀
12	al**	23.04.16	기본 검정 블레이저가 없어서 구매해봤습니다 길이는 엉덩이 덮는 기장감이고 팔은 딱 떨어지는 것 같아요 어깨라인 잡아줘서 이쁘고요 와이드하게 입울때 다 잘 어울리는 것 같아요! 아주 만
13	mi**	23.04.02	어깨가 작아이고 넓은 편이라 너무 오버핏이면 어깨가 태평양갈을 것 같고, 또 너무 딱 맞으면 너무 차려입은 느낌이 날 것 같아서 약간 오버핏인 자켓을 찾고 있었는데요! 딱 이 자켓이 그런 느
14	iu**	23.11.01	키 작고 마른 체형인 분들은 이 자켓이 썩입니다... 당장 사세요... 몇년 전부터 살짝 오버핏인 자켓 사고싶어서 여기저기 사봤는데 아빠청장 빌러입은 것마냥 어깨나 품이 다 커서 반쯤만 5-6벌
15	sh**	23.10.26	울산사는데 배송빨랐어요 162에 55고 s로 주문했습니다 핏은 원하던 정사이즈 핏이고 소매가 손목끝부분까지 오더라구요 조금더 긴게 좋았지만...그래도 기장은 엉덩이 절반은 덮었고 안감색
16	ch**	22.05.02	다른 라인의 자켓을 사고 가격이 좀 더 싸지만 비슷하고 색이 이쪽이 더 마음에 들어서 샀는데 써서 그런지 밝은색이라 그런지 실타가 좀 나요ㅠㅜ 탄탄하지 못해서 반듯한 느낌이 없어요. 그
17	kh**	23.03.23	무난한 자켓으로는 진짜 대박적 아이템입니다 목이 짧고 가슴이 커서 자켓 셔츠류가 정말 안 어울리는 사람인데 이런 찰떡같이 잘 받아요 캐주얼하게 흰 티에 청바지 입어도 되고 안에 원피스
18	w5**	23.09.27	리뷰보고 사는편입니다 오버핏은 싫고 어깨는 맞으며 약간 낙낙한 핏을 원했는데 딱 괜찮구요 어깨는 약간 큼니다 길이가 가장 맘에 들고 두께도 지금 부터 늦가을까지 입기에 좋을것 같
19	kk**	22.04.23	일단 4.21일 밤늦게 주문해서 4.23일 낮에 받았습니다. 퀄리티는 가격값.합니다. 네. 핏도 정사이즈 핏이구요. 리뷰보다는 실밥상태 좋았습니다. 구김도 조금?있었고 냄새는 아무냄새도 안났습
20	hw**	22.04.15	색은 화면과 비슷하구요!다만 등쪽이 타이트한 느낌이 있어요 ㅠㅠ그래서 단추 채운것보단 문게 예쁘더라구요 그리고 생각보다 알지않아서 지금 입기 딱 좋습니다.구김이 잘 생길 수 있어서 조
21	so**	22.09.18	날씨가 갑자기 더워져서 거의 입지 못하고 들고만 있었지만 그래도 핏은 원하던 핏이어서 그냥 그냥 만족합니다~ 그리고 원하던 재질이 아니긴 했지만 핏도 핏이고 급하기도 했고 그래서 그
22	s1**	22.05.06	상체에 살집이 조금 있는 체형이라 핏이 예쁘게 떨어지지 않을까 조금의 걱정이 있었는데 생각했던것보다 훨씬 낙낙한 핏이고 어디에나 믹스매치해도 잘 어울리는 깔끔한 디자인이라 완전 만
23	ye**	22.04.19	생각한 것보다 사이즈가 더 큰 것 같아서 놀랐어요. 근대 그 덕에 핏이 더 예쁜 것 같기도 하네요. 아쉬운 건 가슴쪽에 있는 관상을 주머니(?) 마감이랑 전체적으로 먼지가 붙기 쉬운 재질?이라고
24	dm**	21.11.02	이번에 데죽에서 산 것 중에 가장 마음에 들어요! 우선 핏이 예쁘고 질도 괜찮아요. 다른 곳에서 2배정도 가격의 옷을 샀었는데 더 두껍고 질이 좋은 편이긴 했지만 핏이 너무 안예뻐서 반품했
25	kj**	21.08.12	어깨패드 있는듯 없는듯 하고 전체적으로 가격에 비해 만족하는 핏.디자인인가같아요 아쉬운건 어깨가 좀 더 탄탄했으면 하거나 소매부분이 저렇게 브이자로 마감된거나 가슴부분에 비스듬하

리뷰 정보에는 리뷰어 이름, 리뷰 날짜, 그리고 리뷰 텍스트가 포함되어 있습니다.

같은 방식으로 리뷰 정보도 **pandas**라이브러리를 사용하였고, 최종적으로 **CSV** 파일로

저장하였습니다.

결과물은 간결한 형태로 정리되어 있어, 이후 추가 분석이나 시각화 등에 용이하게 사용할 수 있습니다.