

크롤링 - 더페이스샵

1. 소속 브랜드 URL 추출

(코드)

3.1.더페이스샵 브랜드 접근

```
limit_page = 0
brand_url_list = []
while limit_page < 1:
    try:
        thefaceshop_url = "https://www.naturecollection.com/mall/index.jsp"
        driver.get(url = thefaceshop_url)
        sleep(random.randint(3, 8))

        try:
            ### 브랜드 XPATH
            brand_XPATH = driver.find_elements(By.XPATH, "/html/body/div[2]/div/section/div[2]/ul/li/a")
            sleep(random.randint(1, 5))
            print("브랜드 개수", len(brand_XPATH)-2)

            if len(brand_XPATH) > 1:
                limit_page = limit_page + 1

            for idx in range(1,len(brand_XPATH)-1):
                brand_url = brand_XPATH[idx].get_attribute('onclick')
                brand_url = brand_url[10:38]
                brand_url_list.append(brand_url)
            except:
                limit_page = limit_page + 1
            except:
                limit_page = limit_page + 1

        print("결과 확인" , brand_url_list)
```

브랜드 개수 11
결과 확인 ['/mall/brand.jsp?cate_seq=662', '/mall/brand.jsp?cate_seq=671', '/mall/brand.jsp?cate_seq=663', '/mall/brand.jsp?cate_seq=670', '/mall/brand.jsp?cate_seq=672', '/mall/brand.jsp?cate_seq=669', '/mall/brand.jsp?cate_seq=668', '/mall/brand.jsp?cate_seq=665', '/mall/brand.jsp?cate_seq=666', '/mall/brand.jsp?cate_seq=667', '/mall/brand.jsp?cate_seq=664']

5.파일 저장

```
brand_url_df = pd.DataFrame([brand_url_list])
brand_url_df = brand_url_df.transpose()
brand_url_df.columns = ['url']
print(brand_url_df)

brand_url_df.to_excel("230103_더페이스샵소속브랜드URL.xlsx")
```

(결과물)

	A	T	B	T
1			url	
2	0		/mall/brand.jsp?cate_seq=662	
3	1		/mall/brand.jsp?cate_seq=671	
4	2		/mall/brand.jsp?cate_seq=663	
5	3		/mall/brand.jsp?cate_seq=670	
6	4		/mall/brand.jsp?cate_seq=672	
7	5		/mall/brand.jsp?cate_seq=669	
8	6		/mall/brand.jsp?cate_seq=668	
9	7		/mall/brand.jsp?cate_seq=665	
10	8		/mall/brand.jsp?cate_seq=666	
11	9		/mall/brand.jsp?cate_seq=667	
12	10		/mall/brand.jsp?cate_seq=664	

(코드)

(코드)

```
product_num_df = pd.DataFrame([product_num_list])
product_num_df = product_num_df.transpose()
product_num_df.columns = ['url']
print(product_num_df)
```

```

      url
0  AF015664
1  AF015652
2  AF015635
3  AF015634
4  AF015632
..      ...
226 AF014800
227 AF014750
228 AF014749
229 AF014401
230 AF014399

[231 rows x 1 columns]
```

```
product_num_df.to_excel("230104_제품번호.xlsx")
```

(결과물)

230103_더페이스샵크롤링.ipynb

230104_제품번호.xlsx

230104_제품번호.xlsx

	A	T	B	T
1			url	
2	0		AF015664	
3	1		AF015652	
4	2		AF015635	
5	3		AF015634	
6	4		AF015632	
7	5		AF015626	
8	6		AF015611	
9	7		AF015602	
10	8		AF015601	
11	9		AF015590	
12	10		AF015589	
13	11		AF015572	
14	12		AF015652	
15	13		AF015634	
16	14		AF015611	
17	15		AF015601	
18	16		AF015589	
19	17		AF015565	
20	18		AF015554	
21	19		AF015553	
22	20		AF015536	
23	21		AF015452	
24	22		AF015446	
25	23		AF015445	

230103_더페이스샵크롤링.ipynb

230104_제품번호.xlsx

230104_제품번호.xlsx

	A	T	B	T
208	206		AF011265	
209	207		AC007043	
210	208		AC006820	
211	209		AC006819	
212	210		AC006818	
213	211		AC006817	
214	212		AC005961	
215	213		AC005959	
216	214		AC000177	
217	215		AC000176	
218	216		AC003924	
219	217		AC003923	
220	218		AC003847	
221	219		AC003845	
222	220		AC003844	
223	221		AC003843	
224	222		AF015406	
225	223		AF015405	
226	224		AF015404	
227	225		AF015403	
228	226		AF014800	
229	227		AF014750	
230	228		AF014749	
231	229		AF014401	
232	230		AF014399	

3. 제품 상세정보 추출

(코드)

3.3. 상세 페이지 접속

```
product_url_list = []
category_list = []
brand_name_list = []
product_name_list = []
product_img_list = []
origin_price_list = []
sale_price_list = []
# sale_time_list = []
product_detail_img_list = []
## 구매정보 클릭 후, 수집되는 리스트
ml_info_list = []
REC_skin_type_list = []
use_possible_list = []
how_to_use_list = []
oem_list = []
ingred_list = []
functional_test_list = []

plen = len(product_num_df);
for idx in range(plen):
    print(idx+1, "=====")
    product_url = "https://www.naturecollection.com/mall/product/product-view.jsp?dpid=" + product_num_df['url'][idx]

    try:
        driver.get(url = product_url)
        ## 해당 URL 주소
        product_url_list.append(product_url)
        sleep(random.randint(3, 7))

    try:
```

(결과)

231019_G0444																															
--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

4. 제품 리뷰 정보 추출

(코드)

3.4. 리뷰 페이지 접속

```
product_url_list = []
user_ID_list = []
user_spec_list = []
rating_list = []
review_date_list = []
review_cotents_list = []
review_img_list = []

plen = len(product_num_df)
for idx in range(4):#plen):
    print(idx+1, "=====")
    product_url = "https://www.naturecollection.com/mall/product/product-view.jsp?dpid=" + product_num_df['url'][idx]

    try:
        driver.get(url = product_url)
        ## 해당 URL 접속
        product_url_list.append(product_url)
        sleep(random.randint(3, 7))

    try:
        ## 클릭(리뷰)
        review_button_XPATH = driver.find_elements(By.XPATH, "/html/body/div[2]/div[2]/section/div/div[5]/ul/li[3]/a")
        review_button_XPATH[0].click()
        sleep(random.randint(2,5))

    try:
        ## 사용자 아이디
        user_ID_XPATH = driver.find_elements(By.XPATH, "/html/body/div[2]/div[2]/section/div/div[10]/div[2]/ul/li/div[1]/div[3]/p[1]/span[1]")
```

(결과)

A	B	C	D	E	F	G	H	I	J	K	L	M	N
	url	user_ID	user_spec	rating	review_date	review_cotents							
0	https://www	NONE	NONE	NONE	NONE	NONE							
1	https://www	NONE	NONE	NONE	NONE	NONE							
2	https://www	KA_1*****	30대	5	23.01.04	너무 좋아요잘 쓸게요~감사합니다나한테 잘 맞아요~지인에게도 강추하고싶어요이 제							
3	https://www	KA_1*****	30대	5	23.01.07	겨울철 닥터벨머 시카 리커버리크림을 애용하는데 저렴하게 잘샀네요. 덤으로 어드밴스							
4	https://www	KA_1*****	30대	5	23.01.03	닥터벨머는 순하지만 기능도 좋아서 제가 지인들한테 추천하는 제품인데요!^^토너 바							
5	https://www	11ya*****	30대	4	22.12.09	닥터벨머 어드밴스드 시카 팡타이트 앰플이 처음 나왔을 때 써 보고 좋아 꾸준히 쓰다							