



Vegan? 너도 할 수 있어

1. 기획 의도

글로벌 환경 문제가 대두되고, 동물권에 대한 인식이 높아짐에 따라 비건 생활을 시도하려는 인구가 점차 늘고 있다. 채식의 단계별로 종류가 다양한 데에 반해, 대부분의 비건 어플은 완전한 채식을 하는 사람들의 기준에 맞추어져 있어, Flexitarian(경우에 따라 육식도 하는 경우)이나 Lacto Vegetarian(유제품은 먹는 경우), Ovo Vegetarian(동물의 알을 먹는 경우) 등 다양한 비건 생활인들의 수요에 맞춰진다고 보기는 어렵다고 느꼈다.

그래서 우리 팀은 다양한 이유로 비건을 시작하려고 하는 사람들에게 필요한 서비스들을 고민해 보고 구현하려고 한다. 비건을 처음 접하는 사람도, 비건을 지속하던 사람도 누구든 비건에 쉽게 도전하고, 비건을 생활화 및 습관하여 건강한 습관을 이어갈 수 있도록 도와주는 비건 생활 보조 서비스를 목표로 프로젝트를 기획했다.

비건을 시도하려고 하는 사람들에게 또는 비건을 습관화하기 위해 노력하는 사람들에게 필요한 서비스들에는 무엇이 있을지 고민을 해보며 우리는 다음과 같은 서비스를 구현하고자 한다.

1. 비건 단계별 섭취 가능 식품 분류 및 추천 서비스
2. 동식물성 화장품 분류 및 추천 서비스
3. 비건 식품 및 화장품 후기 모음집
4. 채식 가능 식당 지도 시각화 서비스
5. 비건생활화 어플 서비스 기획

기획서에 들어가기 앞서, 본 기획서 내의 비건은 완전 채식을 하는 비건(Vegan)이 아닌 비거니즘(Veganism)을 실현하려는 사람을 일컫는 표현임을 미리 밝혀두며, 완전 채식을 하는 경우 영어로 Vegan을 병기하였다.

2. 역할 및 일정 분배

1. 역할 분배

이름	역할
성지영(팀 리더)	- 일정 관리 및 프로젝트 관리 - 올리브영 데이터 크롤링 - 데이터 전처리 - 서비스 기획 및 와이어프레임 구성 - 지도 데이터 크롤링
박정호	-마켓컬리 데이터 크롤링 - 데이터 전처리 - 서비스 기획 - 리뷰 데이터 크롤링, 전처리 및 시각화
변재윤	- 올리브영 데이터 크롤링 - 데이터 전처리 - 서비스 기획 - 기존 서비스 리서치 - 리뷰 데이터 크롤링, 전처리 및 시각화
이혜원	- 올리브영 및 마켓컬리 데이터 크롤링 - 데이터 전처리 - 분류 시스템 구현 - 추천 시스템 구현 - 데이터 시각화 - 리뷰 데이터 크롤링 - 지도 데이터 크롤링, 전처리 및 시각화

2. 일정 분배

스케줄

Aa 이름	날짜
-------	----

Aa 이름	📅 날짜
🍌 주제 선정	@2023년 2월 10일 → 2023년 2월 13일
📊 데이터 크롤링	@2023년 2월 13일 → 2023년 2월 16일
📄 데이터 전처리	@2023년 2월 15일 → 2023년 2월 22일
1️⃣ 1차 멘토링	@2023년 2월 18일
📁 분류 시스템 구현	@2023년 2월 20일 → 2023년 3월 7일
👤 추천시스템 구현	@2023년 3월 2일 → 2023년 3월 10일
2️⃣ 2차 멘토링	@2023년 3월 4일
🔥 서비스 기획	@2023년 3월 6일 → 2023년 3월 10일
🌞 데이터 시각화	@2023년 3월 8일 → 2023년 3월 10일
🥗 채식 식당 데이터 크롤링	@2023년 3월 13일 → 2023년 3월 15일
📁 리뷰 크롤링	@2023년 3월 13일 → 2023년 3월 15일
📄 채식 식당 데이터 전처리 및 시각화	@2023년 3월 15일 → 2023년 3월 17일
📁 리뷰 전처리 및 감성분석	@2023년 3월 15일 → 2023년 3월 17일
3️⃣ 3차 멘토링	@2023년 3월 18일
📄 마무리, 발표 준비	@2023년 3월 20일 → 2023년 3월 28일
📢 발표	@2023년 3월 29일

3. 사용한 도구

- Visual Studio Code(.ipynb)
- Python 3
 - Pandas
 - Numpy
 - Matplotlib(pyplot)
 - Beautiful Soup
 - Selenium
 - WordCloud
 - Sklearn
 - pytesseract
 - folium
- Excel(csv, xlsx)
- Clipchamp
- Adobe Express

4. 사전 리서치 및 작업

a. 수집 페이지 선정

‘비건 생활’을 위한 서비스를 기획하기 위해 식품과 화장품, 두 가지 제품군에서 데이터를 수집하기로 결정했다. 생활용품이나 옷과 같이 ‘생활’을

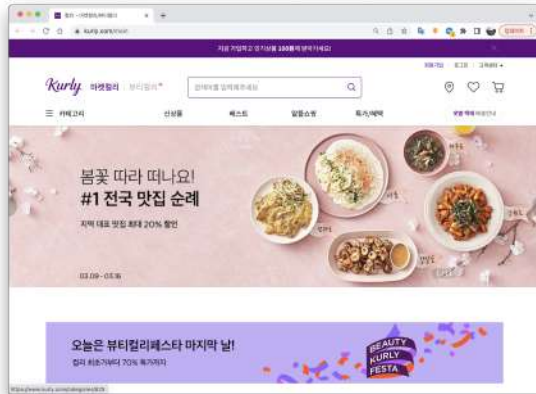
이루는 다양한 제품군이 존재하지만, 앞의 두 가지가 직접 먹고 바르는 등 성분에 대한 분석이 유효한 제품군들이라는 점과 비건 제품의 기준이 명확하다는 점에서 둘에 한정하여 분석을 진행하기로 결정하였다. 11번가, 쿠팡, 네이버 쇼핑 등 다양한 페이지에서 제품을 크롤링 해 오려 시도해 보았으나,

한 쇼핑몰 안에서도 성분이 이미지와 텍스트 등 일정하게 이루어져 있지 않고, 텍스트로만 이루어져 있어도 위치가 가장 위에 있거나 맨 아래쪽에

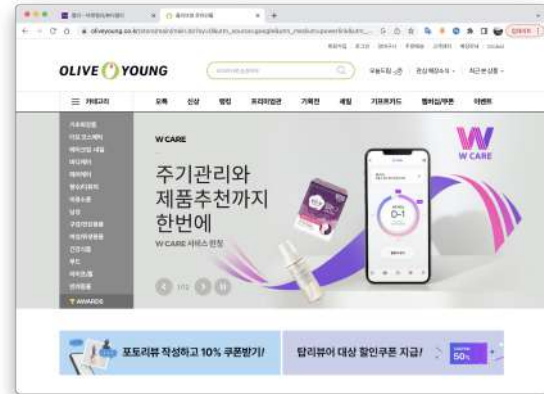
존재하는 등 일정하지 않아 어느 페이지에서 데이터를 수집할지 결정하는 데 어려움을 겪었다.

최종적으로 우리는 식품은 마켓컬리, 화장품은 올리브영에서 크롤링 해 오기로 했다. 그 이유로는 첫 번째, 마켓컬리는 성분표가 모두 이미 지화되어

있고, 올리브영은 성분 위치가 페이지에서 일정한 곳에 있어 크롤링하기에 상대적으로 수월했기 때문이다. 두 번째는 마켓컬리와 올리브영이 각각의 제품군의 상품을 우리가 크롤링 하려는 최소 개수(3000개) 이상으로 가지고 있고, 각각의 제품군의 구매에 있어 영향력 있는 페이지들이라고 판단했기 때문이다.



마켓 컬리 홈페이지



올리브영 홈페이지

마켓컬리의 상품들 중에서도 ‘채소’, ‘과일, 견과, 쌀’, ‘수산물, 해산물, 건어물’, ‘정육, 계란’에 해당하는 상품들은 성분 표기가 의미 없는 제품들이라고 판단하여 제외하였고, ‘와인’, ‘전통주’의 경우 주류이기 때문에 성인 인증이 필요해 크롤링이 번거로워 제외하였다. 그 외 ‘생활용품, 리빙, 캠핑’, ‘스킨케어, 메이크업’ 등 식품에 해당하지 않는 제품군들도 모두 제외하였다. 그렇게 남은 ‘국·반찬·메인요리’, ‘샐러드·간편식’, ‘면·양념·오일’, ‘생수·음료·우유·커피’, ‘간식·과자·떡’, ‘베이커리·치즈·델리’의 여섯 가지 제품군에서 각각 500개 이상의 제품을 크롤링 하여 총 3000개 이상의 식품 이름과 제조사, 알레르기 성분 정보, 성분표 이미지를 크롤링 하였다.

올리브영의 경우 ‘더모 코스메틱’이라는 카테고리의 제품군이 대부분 다른 카테고리의 제품군들과 겹친다는 것을 확인하여 제외하였고, ‘미용소품’이라는 카테고리의 제품군은 화장품이 아니라 주변 기기에 가까워 제외하였다. 그렇게 10가지 카테고리에서 각각 300개 이상의 제품을 크롤링 하여 총 3000개 이상의 화장품의 이름, 제조사, 성분, 비건 여부를 크롤링 하였다.

b. 성분리스트 수집

i. 성분 리서치 및 크롤링

성분을 기준으로 채식의 단계를 나누는 작업을 진행하기 때문에 식품과 화장품의 성분에 대한 자료 조사가 많이 필요했다. 수입식품정보마루(<https://impfood.mfds.go.kr/>), 대한화장품협회(<https://kcia.or.kr/cid/main/>), 식품안전나라(<https://www.foodsafetykorea.go.kr/main.do>) 등 식품과 화장품에 대한 성분들을 정보를 제공하고 있는 사이트들이 존재하였다. 엑셀로 파일을 제공하는 곳도 있었고, 크롤링을 통해 정보를 수집해야 하는 사이트들도 있었다.



수입식품정보마루



대한화장품협회



식품안전나라

크롤링으로 정보를 수집해야 하는 사이트로는 식품안전나라가 있었다.

식물과 동물 탭이 따로 있어 두 번에 걸쳐 크롤링을 진행하였다. 마켓컬리나 올리브영과 달리 원하는 정보가 테이블 형태로 저장되어 있었기 때문에 테이블을 한 번에 크롤링 하는 코드를 작성하였다.

```
url='https://www.foodsafetykorea.go.kr/portal/safefoodlife/foodMaterial/foodMaterialDB.do'

driver=webdriver.Chrome()
driver.get(url)
act=ActionChains(driver)
html=driver.page_source
soup=BeautifulSoup(html, 'html.parser')

# [식물] 클릭
driver.find_elements(By.CSS_SELECTOR, '#menuTab3')[0].click()

# 테이블 크롤링
html=driver.page_source
soup=BeautifulSoup(html, 'html.parser')
data=soup.find('table')
table=parser_functions.make2d(data)

# 데이터프레임
df=pd.DataFrame(data=table[1:], columns=table[0])

# 다음 페이지
driver.find_elements(By.CSS_SELECTOR, '#tab3 > div.board-footer > div > ul > li:nth-child(7) > a')[0].click()

page=1
while 1:
    try:
        page+=1
        # 테이블 크롤링
        html=driver.page_source
        soup=BeautifulSoup(html, 'html.parser')
        data=soup.find('table')
        table=parser_functions.make2d(data)

        # 데이터프레임
        df2=pd.DataFrame(data=table[1:], columns=table[0])
        df=pd.concat([df, df2])

        # 다음 페이지
        driver.find_elements(By.CSS_SELECTOR, '#tab3 > div.board-footer > div > ul > li:nth-child(7) > a')[0].click()

    except:
        break

    if page%100==0:
        print(page)
```

이 코드를 사용하여 동물 탭을 클릭하여 크롤링 하는 작업도 진행하였다.

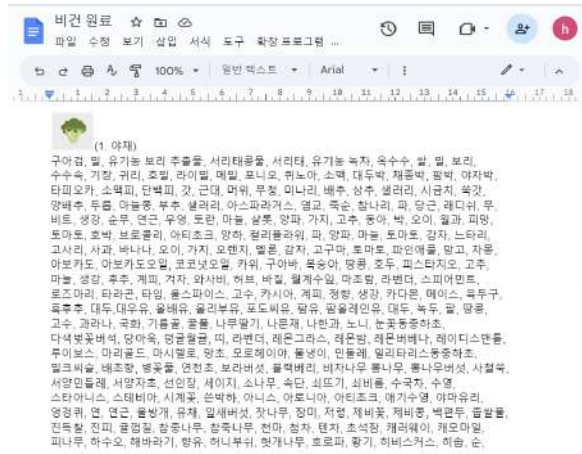
	원재료	품목번호	이명 또는 영명	학명 또는 특성	생약명	가능	제한	사용부위	사용조건	비고
0	Abiu	-	Yellow star apple, Caimito, Cairno, Luma	Pouteria caimito Radlk / Lucuma caimito Roem. ...	-	O	X	열매	-	-
1	Abobra tenuifolia	-	-	Abobra tenuifolia	-	O	X	열매	-	-
2	Abutilon	-	Painted Indian Mallow	Abutilon pictum	-	O	X	꽃	-	-
3	Abyssinian cabbage	-	Abyssinian Mustard	Brassica carinata	-	O	X	잎	-	-
4	Acanthus-leaved Thistle	-	Golden Thistle	Carlina acanthifolia All.	-	O	X	꽃	-	-
...
5	Alpine heuchera	-	-	Heuchera glabra	-	O	X	잎	-	-
6	Alpine spring beauty	-	-	Claytonia megarhiza	-	O	X	뿌리, 잎, 꽃	-	-
7	Alstroemeria revoluta	-	-	Alstroemeria revoluta	-	O	X	뿌리	-	-
8	Alstroemeria spectabilis	-	-	Alstroemeria spectabilis	-	O	X	뿌리	-	-
9	Alstroemeria versicolor	-	-	Alstroemeria versicolor	-	O	X	뿌리	-	-

70 rows x 10 columns

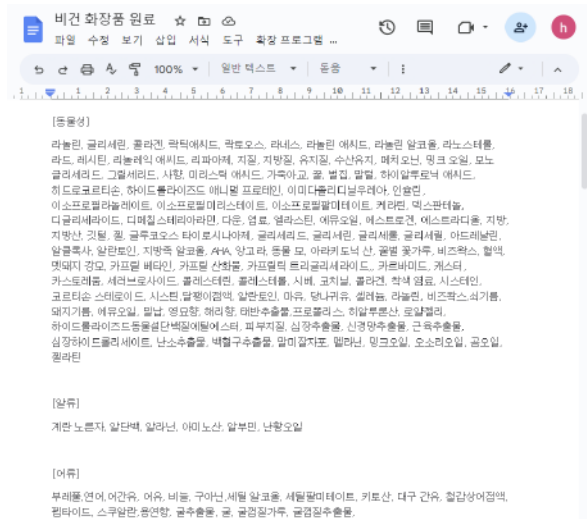
식품안전나라 크롤링 결과 데이터셋

수집한 데이터 중 [원재료, 이명 또는 영명, 생약명, 사용부위]를 추출하여 text파일로 저장하였다.

그 외에도 구글링을 통해 다양한 화장품과 식품의 성분들에 대한 정보를 모았고, Google Docs를 통해 팀원들과 함께 정리하였다.



식품 성분 정리



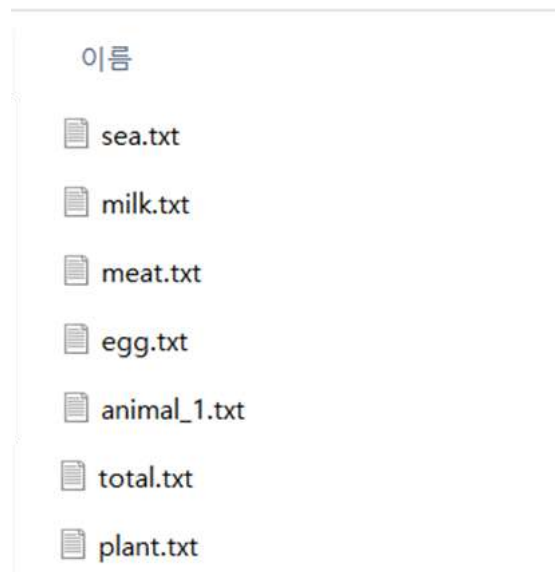
화장품 성분 정리

ii. 성분 리스트 통합 및 전처리

다운로드, 구글링, 크롤링 등을 통해 수집한 식품들을 분류에 사용할 리스트로 만들기 위해서는 재료를 분류해주는 작업이 필요하였다. 따라서 각 성분을 육류, 해산물, 유제품, 동물의 알, 식물, 식품첨가물 등으로 나눠주었다.

식품안전_수입식품				
파일 수정 보기 삽입 서식 데이터 도구 ...				
100% W % .0 .00 123 기본값... 10				
A1	원재료			
	A	B	C	D
1	원재료	이명 또는 영명	생약명	분류
3	Adelomelon ancilla	-	-	해산물
4	Alaska skate	-	-	해산물
5	Alewife	Gaspereau	-	해산물
6	American harvestfish	-	-	해산물
7	Argentine shortfin squid	-	-	해산물
8	Atlantic blue marlin	Atlantic blue marlin	-	해산물
9	Atlantic hagfish	-	-	해산물
10	Atlantic halibut	-	-	해산물
11	Atlantic menhaden	Fatback, Bugfish, Bunker, Mossbunker, Pogy	-	해산물
12	Atlantic sardine	Canadian sardine, Herring	-	해산물
13	Atlantic wolffish	Striped wolffish	-	해산물
14	Backwater hard clam	-	-	해산물
15	Beaked redfish	-	-	해산물
16	Bengal tongue sole	-	-	해산물
17	Bigeye croaker	Bigeye croaker	-	해산물
18	Bighead catfish	-	-	해산물
19	Black scabbardfish	-	-	해산물
20	Blue ling	-	-	해산물
21	Bridled grouper	-	-	해산물
22	Brown crab	edible crab	-	해산물
23	Brown tiger prawn	Common tiger prawn	-	해산물
24	Catla	-	-	해산물
25	Commerson's anchovy	-	-	해산물

식품 성분들은 정리한 각각의 리스트들을 따로 txt파일로 저장하였고, 화장품 성분들은 동물성 리스트와 식물성 리스트로 나눠서 txt파일로 저장하였다.



5. 서비스 구현

a. 비건 단계별 식품 분류 및 추천 시스템 구현

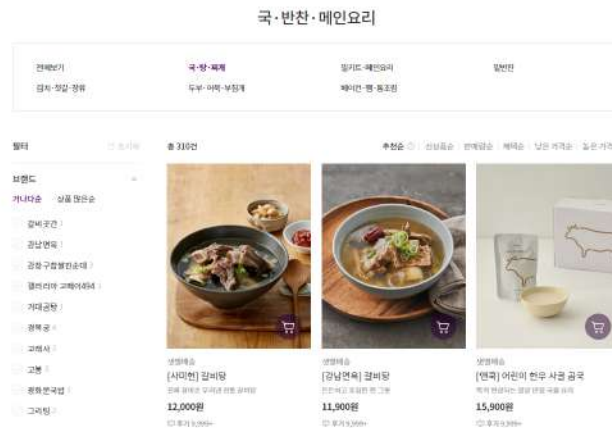
i. 데이터 수집

마켓컬리 사이트에서 자체적으로 분류한 카테고리들 중 국·반찬·메인요리, 샐러드·간편식, 면·양념·오일, 생수·음료·우유·커피, 간식·과자·떡, 베이커리·치즈·델리 카테고리를 선택하여 크롤링을 진행하기로 하였다.

카테고리	신상품	베스트
채소	국·탕·찌개	
과일·견과·쌀	밀키트·메인요리	
수산·해산·건어물	밑반찬	
정육·계란	김치·젓갈·장류	
국·반찬·메인요리	두부·어묵·부침개	
샌드위치·간편식	베이컨·햄·통조림	
면·왕념·오밀		
생수·음료·우유·커피		
간식·과자·떡		
베이커리·치즈·얼리		
건강식품		
와인		
전통주		
생활용품·리빙·캠핑		
스킨케어·메이크업		
헤어·바디·구강		
주방용품		
가전제품		
반려동물		
베이비·키즈·완구		

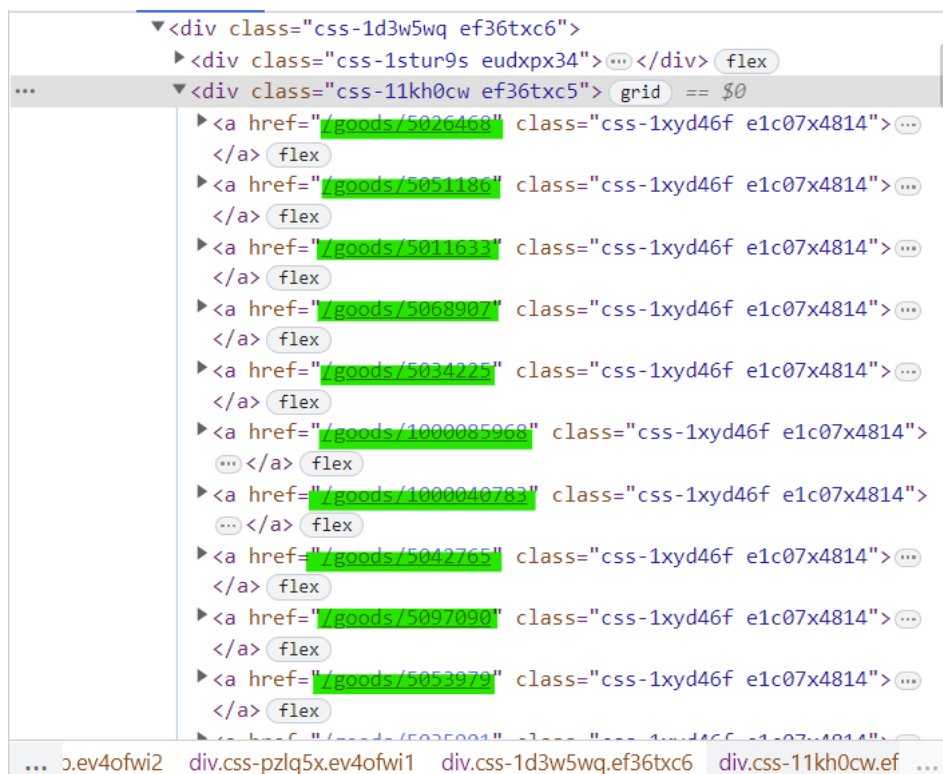
마켓컬리 카테고리 종류

각 카테고리마다 다시 소카테고리로 나뉘는데, 소카테고리로 들어가게 되면 한 페이지 당 96개의 상품을 보여주고 있고, 다음 페이지로 넘어가는 숫자 버튼이 존재했다.



마켓컬리 소카테고리 페이지

페이지의 화면 구성을 살펴봤을 때 각 상품마다 상품에 대한 고유번호가 존재했고, 이를 활용하여 상품의 상세페이지로 이동할 수 있다는 것을 확인할 수 있었다.



마켓컬리 소카테고리페이지 화면구성

따라서 각 상품의 고유 상품번호를 받아와 product_url이라는 리스트에 저장하는 크롤링 코드를 먼저 작성하였다.

```
for url in url_lst:
    driver=webdriver.Chrome()
    driver.get(url)
    act=ActionChains(driver)

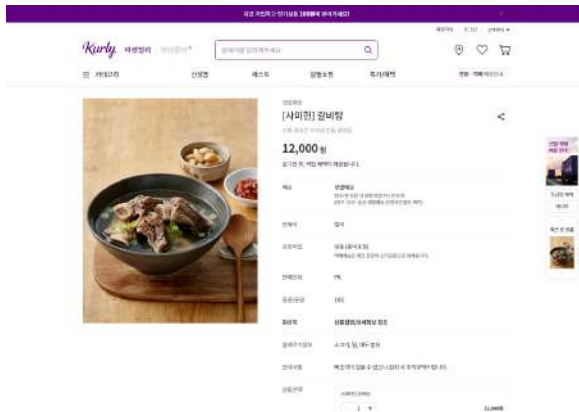
    html=driver.page_source
    soup=BeautifulSoup(html, 'html.parser')

    a=soup.select('div.css-11kh0cw.ef36txc5 > a')
```



```
for i in range(len(a)):
    product_url.append(a[i].attrs['href'])
```

상품을 클릭하여 상품페이지로 들어가면, 상품명, 가격, 알레르기 정보 등이 존재했다. 우리가 필요로 하는 성분 정보 이미지 형태로 존재했다.



마켓컬리 상품 상세페이지



마켓컬리 상품 상세페이지 성분이미지

원하는 정보들의 위치를 파악해 봤을 때, 이전에 크롤링 해온 상품 고유번호를 이용하여 url을 새로 지정하여 상품페이지로 이동 후, 상품의 이름과 알레르기 정보, 성분 이미지를 크롤링 한 후 창을 닫고 다음 상품 고유번호를 이용해 다른 상품페이지로 이동하는 과정을 반복해 주는 코드를 작성해주었다.

성분 이미지의 경우 새로운 url 주소로 저장되어있기 때문에 우선 url을 저장해주었다.

```
for idx,product in enumerate(product_url):
    url='https://www.kurly.com'+product
    driver=webdriver.Chrome()
    driver.get(url)
    act=ActionChains(driver)
    html=driver.page_source
    soup=BeautifulSoup(html,'html.parser')
    allergy=''

    try:
        #브랜드
        item = soup.select('div > h1.css-1f2zq3n.ezpe9l11')[0].text
        item_lst.append(item)

        # 성분이미지출
        details = driver.find_elements(By.ID, 'detail')
        detail_img =details[0].find_element(By.TAG_NAME, 'img').get_attribute('src')
        img_lst.append(detail_img)

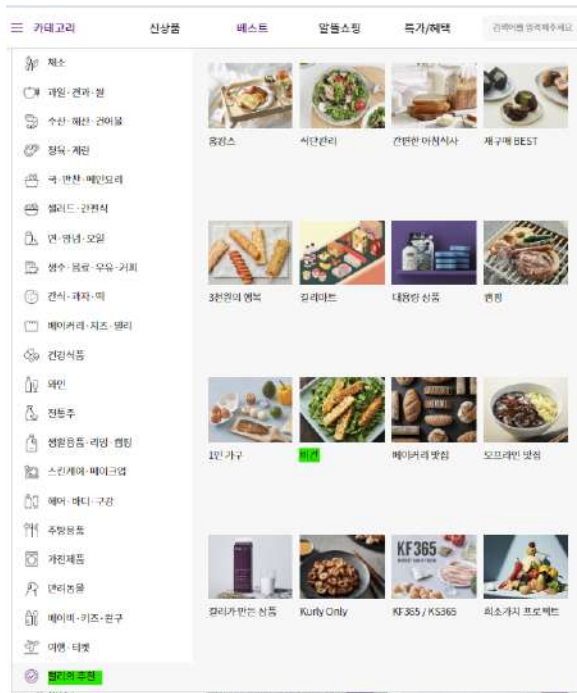
        # 알레르기 정보
        allergy = soup.select('dl > dd > p')[8].text
        alr_lst.append(allergy)

    except:
        alr_lst.append(allergy)
        idx_lst.append(idx)

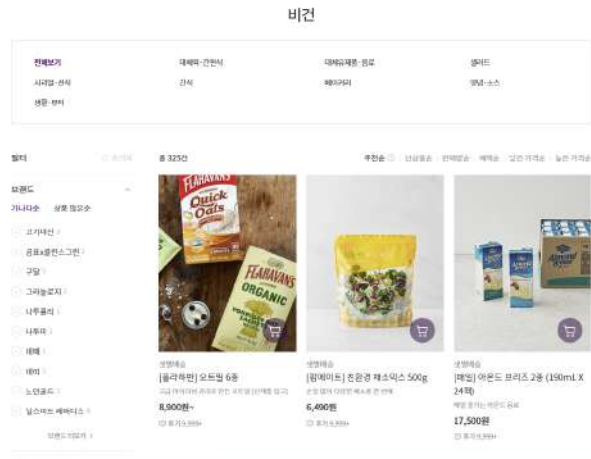
    driver.close()
```

이와 같은 작업을 소카테고리마다 반복하여 주었다.

마켓컬리가 자체적으로 분류해둔 카테고리 중 '컬리의 추천'에는 '비건'이라는 탭이 따로 존재한다.



마켓컬리 카테고리 비건 탭



마켓컬리 비건 카테고리 상품 예시

우리가 이전에 크롤링 한 상품들 중 비건인 상품은 무엇이 있는지 확인하기 위해 상품의 이름이 필요했다. 따라서 상품 상세페이지로는 이동하지 않고 비건 카테고리 페이지에서 상품의 이름을 크롤링 하는 작업을 진행하였다.

```
for page in range(10):
    html=driver.page_source
    soup=BeautifulSoup(html, 'html.parser')

    try:
        for i in range(96):
            vegan_lst.append(soup.select('div > a > div > span.css-1d9y2r1.e1c07x488')[1].text)

            driver.find_elements(By.CSS_SELECTOR, '#container > div.css-pz1q5x.ev40fw11 > div.css-1d3w5wq.ef36txc6 > div.css-rdz8z7.e821nfz')
    except:
        break
```

수집한 비건 데이터 중에 이전에 수집한 마켓컬리 데이터에 존재하는 상품들에 한해서 합하여 최종 데이터 셋을 완성해 csv파일로 저장하였다.

category	item	allergy	img	vegan
음료/우유/커피/차	[콜린스그린] 원데이 클렌즈 (350mL X 5개)	없음	https://img-cf.kurly.com/shop/data/goodsview/2...	0
음료/우유/커피/차	[콜린스그린] 더 오렌지 1000mL	없음	https://img-cf.kurly.com/shop/data/goodsview/2...	0
음료/우유/커피/차	[존시몬] 오렌지 주스 2L	없음	https://img-cf.kurly.com/shop/data/goodsview/2...	0
음료/우유/커피/차	[YOZM] 플레인 그릭요거트 소프트 500g	없음	https://img-cf.kurly.com/shop/data/goodsview/2...	0
음료/우유/커피/차	[풀무원] 아임리얼 700mL 7종	- 본 제품은 토마토를 사용한 제품과 같은 제조 시설에서 제조하고 있습니다.	https://img-cf.kurly.com/shop/data/goodsview/2...	0

마켓컬리 데이터 크롤링 후 데이터프레임

ii. 데이터 전처리

전처리가 필요한 부분을 살펴보았다.

우선 상품 이름을 살펴보면 대괄호(()) 안에 브랜드명이 적혀있는 것을 확인할 수 있다. 이를 brand와 product로 분리해 주기로 했다. 대괄호가 없는 상품은 마켓컬리의 자체 상품이라고 판단하여 브랜드명을 kurly라고 지정해 주었다.

```
for item in item_lst:
    if len(item.split(' '))>1:
```

```

        brand=item.split(' ')[0]
        brand=brand.replace('[', '')
        product=item.split(' ')[1]
        brand_lst.append(brand.strip())
        product_lst.append(product.strip())
    else:
        brand_lst.append('kurly')
        product_lst.append(item.strip())

```

알레르기 성분이 아닌 제조회사를 설명하는 문자들을 찾아 제거해 주었고, 숫자가 포함되어 있다면 제거하는 작업을 했다. 알레르기에 대한 정보가 없는 상품에 대해서는 '없음'으로 통일하여 주었다.

```

result_lst=[]
for text in df['allergy']:
    # [] 내용 제거
    cnt=0
    try:
        while(1):
            start_idx=re.search('[',text).start()
            end_idx=re.search(']',text).end()
            text=text[:start_idx+1]+text[end_idx+1:]
            cnt+=1
            if cnt>4:
                break
    except:
        pass

    result2=''
    for word in text.split('\n'):
        try:
            idx=re.search(']',word).start()
            word=word[idx+1:]
        except:
            pass

        # 기타 처리
        word=word.replace(' ', '')
        word=word.replace(':', ',')
        word=word.replace(' 함유식품', '')
        word=word.replace(' 함유', '')
        word=word.replace('이제품은', '')
        word=word.replace('본제품은', '')

        # 함께제조 삭제
        try:
            start_idx=re.search('을사용',word).start()
            word=word[:start_idx]
        except:
            pass

        try:
            start_idx=re.search('와같은',word).start()
            word=word[:start_idx]
        except:
            pass

        try:
            start_idx=re.search('를사용',word).start()
            word=word[:start_idx]
        except:
            pass

        try:
            start_idx=re.search('과같은',word).start()
            word=word[:start_idx]
        except:
            pass

        if '상품별' in word:
            word=''

    ## ,로 다시 나눠서 숫자 포함되어 있으면 삭제
    result=''
    for ing in word.split(','):
        if hasNumber(ing)==True:
            pass
        else:
            result+=ing+' '

    result2+=result+' '

```

```
result_lst.append(result2)
```

```
for text in result_lst:
    text=re.sub('[^a-zA-Z가-힣,()]\s','',text)
    text=re.sub(' ','',text)

    if len(text)==0:
        text='없음'
    if text[0]==' ':
        text=text[1:]

    if len(text)==0:
        text='없음'
    if text[-1]==' ':
        text=text[:len(text)-1]

    if '별도' in text:
        text='없음'

    text=text.replace(' (공통)', '')

    lst.append(text)
```

이후 데이터를 살펴봤을 때 '혼입가능'이라는 단어가 포함되어 있는 것을 확인할 수 있었다. 또한 소괄호 안에 알레르기를 불러일으킬 수 있는 재료들이 적혀있었기 때문에 괄호는 삭제하되 안에 내용은 살려두는 작업이 필요하다고 판단되었다.

```
for idx,text in enumerate(kurly['allergy']):
    text=text.replace(' 혼입가능', '')
    text=text.replace(' (', ', ')
    text=text.replace(' )', ', ')
    text=re.sub(' ','',text)
    kurly['allergy'][idx]=text
```

이렇게 전처리를 마친 총 3756개의 데이터를 csv파일로 저장하였다.

	category	brand	item	allergy	img	vegan
0	국/반찬/메인 요리	풀무원	국산콩 나뭇잎 3종	국산콩생나뭇잎, 대두, 밀, 쇠고기, 검은콩생나뭇잎, 대두, 밀, 쇠고기, 국산콩와사비나뭇잎, 대두, 밀...	https://img-cl.kurly.com/shop/data/goodsview/2...	0
1	국/반찬/메인 요리	kurly	전통 지방 물떡 (2팩)	밀, 대두, 우유, 계란, 우유, 메밀, 팥, 콩, 고등어, 게, 새우, 돼지고기, 복숭아, 토마토, 아황산염...	https://img-cl.kurly.com/shop/data/goodsview/2...	0
2	국/반찬/메인 요리	My little recipe	단종에서 올라온 전통찜닭 대용량 (3~4인분)	닭고기, 오징어, 된대, 된대, 닭고기, 오징어, 우유, 대두, 콩...	https://img-cl.kurly.com/shop/data/goodsview/2...	0
3	국/반찬/메인 요리	누데이블	메추리알 찰조림 350g	알류, 메추리알, 대두, 밀, 조개류, 굴, 알류, 계란, 우유, 메밀, 팥, 콩, 고등어, 게, 새우, 돼지고...	https://img-cl.kurly.com/shop/data/goodsview/2...	0
4	국/반찬/메인 요리	모두익혔친	소문난 팔초 조별낙지 낙곱새	대두, 밀, 새우, 쇠고기, 조개류, 굴, 난류, 우유, 메밀, 팥, 콩, 고등어, 게, 돼지고기, 복숭아, 토...	https://img-cl.kurly.com/shop/data/goodsview/2...	0
...
3751	간식/과자/떡	청우	마음 버 구워만든 죽염렌디 38g	밀, 대두, 우유, 복숭아	https://img-cl.kurly.com/shop/data/goodsview/2...	0
3752	간식/과자/떡	기도업	스타벅아초콜렛	우유, 호두, 밀, 대두, 대두	https://img-cl.kurly.com/shop/data/goodsview/2...	0
3753	간식/과자/떡	타르두프랑게	트뤼플맛 아소티드 스위트 초콜렛 70g (팩)	밀, 우유, 대두, 보리, 팥, 계란	https://img-cl.kurly.com/shop/data/goodsview/2...	0
3754	간식/과자/떡	kurly	맛다시매 블랙 100g (당량)	없음	https://img-cl.kurly.com/shop/data/goodsview/2...	0
3755	간식/과자/떡	미자면니네 발맛간	발고구마 사나온 찰면	계란, 우유, 견과류, 대두, 밀	https://img-cl.kurly.com/shop/data/goodsview/2...	0

마켓컬리 데이터 전처리 완료 데이터프레임

iii. 이미지 텍스트 추출

마켓컬리의 성분 정보들은 이미지로 저장되어 있었고, 이는 url로 데이터 프레임에 저장해두었다. 따라서 우선 url을 이용해서 이미지를 저장하는 과정이 필요했다. 마켓컬리 데이터 프레임의 인덱스 순으로 이미지를 다운 받아 와 인덱스 번호를 이름으로 이미지를 저장하였다.

```
for idx,url in enumerate(data['img']):
    urllib.request.urlretrieve(url, './img/'+str(idx)+'.jpg')
```




My little recipe 안동에서 올라온 전통찜닭 대용량 (3-4인분)

마켓컬리 성분 이미지 - 원재료 가로



모노키친 바지락 납작 칼국수 (613g)

마켓컬리 성분 이미지 - 원재료 세로

성분이 적혀있는 부분이 성분으로 시작하는 상품, 원료로 시작하는 상품, 원재료라고 가로로 쓰여 있는 상품, 원재료라고 세로로 쓰여 있는 상품, 여러 표가 붙어있는 상품 등 다양했다.

이 모든 상품들을 하나의 코드로 통일하여 성분 텍스트를 추출하는 것에는 문제가 있다고 판단이 들었다. 이를 해결하기 위해 다양한 방법을 고민해 보았다.

1. 성분 리스트를 완성하여 이에 해당하는 단어들만 추출한다.
2. 사진을 규칙에 맞게 분류하는 알고리즘을 짜서 사진을 분류한 뒤 이미지에서 성분 텍스트를 추출한다.
3. 사진을 직접 확인해가며 비슷한 모양을 가진 사진들끼리 사진을 분류한 뒤 이미지에서 성분 텍스트를 추출한다.

성분 리스트를 완성해서 이에 해당하는 단어들만 추출하는 1안의 방법이 가장 좋은 방법이라고 판단하였지만 이는 우리가 가지고 있는 성분 리스트에 해당하는 단어들만 추출할 수 있기 때문에 성분 리스트가 어느 정도 완벽하게 구축된 상태일 경우에 가장 좋은 효과를 볼러한다고 판단했다.

2안의 작업을 하기 위해서는 train 데이터에 대해서는 수작업이 필요하기로 하고, 성분표의 생김새와 구조 등이 비슷하면서도 달라서 컴퓨터로 이를 구현하기는 쉽지 않아 보였다.

따라서 시간이 조금 걸리고 귀찮은 수작업이라고 하더라도 1차적으로 사진을 분류한 뒤 규칙을 찾을 수 있는 곳에서 성분들을 뽑아내고, 이전에 리서치를 통해 수집한 성분 리스트에 추가하여 2차 리스트를 만들어주기로 하였다. 이렇게 만든 2차 리스트를 활용하여 1안의 작업을 거친다면 조금 더 많은 이미지에서 더 다양한 성분들을 추출할 수 있을 것이라고 판단하였다.

- 성분
- 원료
- 원재료-가로
- 원재료-세로
- 특이케이스
- 표불어있는것

마켓컬리 데이터 성분 이미지 분류 기준

마켓컬리 상품 이미지들을 성분, 원료, 원재료가 가로로 적혀져 있는 상품, 원재료가 세로로 붙어져 있는 상품, 특이한 케이스, 성분표가 붙어있는 경우로 나눠서 분류해 주었고, 이 중 성분, 원료, 원재료-가로에서 규칙을 찾을 수 있었다.

pytesseract 라이브러리를 이용해 이미지를 텍스트화 시켜주었고, 성분, 원료, 원재료 등의 말이 위치해있는 index 값을 찾고 중량, 보관, 반품 등 성분에 대한 정보 다음에 등장하는 단어의 index 값을 찾아 그중 문장의 길이가 가장 짧은 길이로 slice 해주었다.

```
error_lst=[]
result_lst=[]
pass_lst=[]
for idx in idx_lst:
    img=Image.open(path+idx)
    result=pytesseract.image_to_string(img, lang='kor')

    result=result.replace(' ','')
    a=re.sub('[\n]','',result)

    try:
        # 원료로 시작하는 부분 찾기
        start_idx=re.search('원료',a).start()

        # 끝나는 부분 찾기
        min_len=10000
        min_idx=0
        word_idx=0
        word_lst=['중량','보관','반품','포장','내용','플리','나용','본제품','유전자','용량','품목',
                 '영양','내리','유통','포징','포함','제조원','면리에틸','구입','식품','소비'] # 계속 추가하기
        while 1:
            # break문
            if word_idx>len(word_lst):
                break

            # 최소글자수 찾기
            try:
                end_idx=re.search(word_lst[word_idx],a).start()
                if end_idx<start_idx:
                    word_len=len(a[start_idx:])
                else:
                    word_len=len(a[start_idx:end_idx])

                if word_len<min_len:
                    min_len=word_len
                    min_idx=word_idx

                word_idx+=1

            except:
                word_idx+=1

        end_idx=re.search(word_lst[min_idx],a).start()
        result_lst.append(a[start_idx:end_idx])
        #print(a[start_idx:end_idx],'\n')

        pass_lst.append(idx)

    except:
        error_lst.append(idx)
```

```
[ '원료명:유기농단풍시럽(0090)',
  '원료명:유기농압착올리브유100%',
  '원료및함량:대592%',
  '원료명및함량:계피(100)',
  '원료명:유기농압착아보카도오일100%*반웅장수:',
  '원료명:포도원액(무수아황산(산화방지제))70%, 와인식초(포도,메타중아황산칼륨)30%',
  '원료명201국산)10%, 곤드래나물(국산)10%, 루지갱이나물(국산)10%,가지(아비몬빠티비그거0,계란,콩기름,고추장,0018009309211',
  '원료명및함량:유기농옥수수(83.8%),정제수,정제소금',
  '원료명:브유100%*',
  '원료및함량:가쓰오부시(4301:',
  '원료명:포도원액70%,와인식초29.95무수아황산산화방지제)*',
  '원료명-박력분,리코타치즈(유청,우유,유크림,정제소금,구연산)229,계란,시금치1196,듀럼밀,정제수,브레드크럼(박력분,정제수,
  '원료명:압착올리브유99.89%,천연바실릴0.2%*',
  '원료명1이미티애나지락살(중국|로[데루에티0006000여|미다따다기해름.밀.토마토:징어.조개류(물,함칫5:',
  ''
```

마켓컬리 성분 데이터 텍스트화 및 전처리 후 결과

이후 성분 데이터에서 필요 없는 부분인 국산, 중국산과 같은 단어들을 stopwords에 저장하고, 원료 및 함량, 원료명과 같은 필요 없는 단어들을 제거한 후 정규 표현식을 거쳐 성분들만 남도록 정리해 주었다. 완성한 성분 리스트에 해당하는 단어들을 추출할 예정이기 때문에 통일하기 위해 hanspell을 이용해 맞춤법 검사를 진행한 후 리스트에 추가해 주었다.

```
ingredient_lst=[]
for result in result_lst:
    if len(result)==0:
        sen=''
    elif '원료및함량' in result:
        sen=result[5:]
    elif '원료명및함량' in result:
        sen=result[6:]
    elif '원료명및' in result:
        sen=result[4:]
    elif '원료명' in result:
        sen=result[3:]

    sen=re.sub('([()/*-| | :.{}%!]',' ',sen)
    sen=re.sub(',+',', ',sen)
    sen=re.sub('[^가-힣,]',', ',sen)

    lst=[]
    if ',' in sen:
        for word in sen.split(','):
            word=word.strip()
            if len(word)>0:
                try:
                    spelled_sent=spell_checker.check(word)
                    spelled_sent=spelled_sent.as_dict()
                    word=spelled_sent['checked']
                    word=word.strip()
                except:
                    word=word.strip()

            if len(word)>0 and word not in stopword:
                lst.append(word)
    else:
        lst.append(sen.strip())
    ingredient_lst.append(lst)
```

같은 방식으로 성분, 원재료-가로에도 적용하여 이미지에서 성분에 해당하는 텍스트를 추출하였다.

	img	성분
0	1012.jpg	[유기농 단풍시럽]
1	1042.jpg	[유기농 압착 올리브유]
2	1057.jpg	[]
3	1073.jpg	[계피]
4	1089.jpg	[유기농 압착 아보카도 오일, '반응 장수]
...
1098	2385.jpg	[연성가공치즈, '자연치즈', '탈지유', '슈크림', '버터밀크파우더', '딤...
1099	2386.jpg	[계란, '초콜릿', '코코아', '코코아 버터', '설탕', '레시틴', '바...
1100	2387.jpg	[계란, '식물성 크림', '정제수', '액상과당', '경화광핵유', '레 세틴...
1101	2388.jpg	[계란, '식물성 크림', '정제수', '액상과당', '경화유', '레 세틴',...
1102	2533.jpg	[찰 보릿가루, '우유', '설탕', '게린쿠넵매실엑기스', '계란]

마켓컬리 성분 데이터 이미지 텍스트추출 결과

이렇게 이미지에서 성분에 해당하는 텍스트를 추출한 성분 리스트와 기존에 리서치를 통해 구성한 1차 성분 리스트를 결합하여 2차 성분 리스트를 만들어주었다.

```
# 규칙 리스트
lst=df['성분'].tolist()
ingredient_lst=[]
for sen in lst:
    sen=re.sub("['\[\]]",',',sen)
    for word in sen.split(','):
        word=word.strip()
        if len(word)>0 and len(word)<15:
            ingredient_lst.append(word)
ingredient_lst=list(set(ingredient_lst))

# 1차 리스트 불러오기
with open("./list/total.txt", "r") as file:
    data = file.readlines()

for word in data:
    word=re.sub('\n','',word)
    ingredient_lst.append(word)
```

이렇게 완성한 성분 리스트를 사용하여 성분 이미지에서 성분 텍스트를 추출하는 작업을 진행하였다.

이미지 전체를 순서대로 불러와 pytesseract를 이용해 텍스트를 추출하고, 정규 표현식을 사용하여 특수문자를 제거하고 숫자 등을 제거하며 필요한 부분들만 남겨주었다. 이후 단어를 하나씩 맞춤법 검사를 진행하고 성분 리스트에 있는 단어들을 저장해 주는 과정을 거쳐 이미지에서 성분 정보를 추출하였다.

```
result_lst=[]
for idx in idx_lst:
    # 텍스트 추출
    img=Image.open(path+idx)
    result=pytesseract.image_to_string(img, lang='kor')

    # 특수문자제거
    result=re.sub('[\(\)\[\]\|:\n·%\{\}\.]','',result)
    result=re.sub('[^가-힣,]','',result)
    result=re.sub(' ','',result)

    idx_ingredient_lst=[]
    for word in result.split(','):
        word=word.strip()
        if len(word)>1:
            # 맞춤법 검사
            try:
                spelled_sent=spell_checker.check(word)
                spelled_sent=spelled_sent.as_dict()
                word=spelled_sent['checked']
            except:
                word=word.strip()

    # 성분리스트에 있는 거 저장
```

```

if word in lst:
    idx_ingredient_lst.append(word)

idx_ingredient_lst=list(set(idx_ingredient_lst))
result_lst.append(idx_ingredient_lst)

```

이렇게 3756개의 이미지 데이터에서 성분에 해당하는 텍스트를 추출하고 전처리를 거쳐서 데이터 프레임을 완성하였다.

	category	brand	item	allergy	ingredient
0	국/반찬/메인 요리	동우원	국산콩 나트 3종	국산콩생나트,대두,밀,쇠고기,검은콩생나트,대두,밀,쇠고기,국산콩와사비나트,대두,밀...	['고등어', '아황산류', '맥밀', '옥수수유 옥수수', '복숭아', '오징어...']
1	국/반찬/메인 요리	kurly	전통 시장 물떡 (2팩)	밀,대두,우유,계란,우유,메밀,땅콩,고등어,게,새우,돼지고기,복숭아,토마토,아황산류...	['조개류', '정제소금', '주정', '맥류', '조기', '소스 소스', '참치...']
2	국/반찬/메인 요리	My little recipe	연등에서 올라온 전통찜닭 대용량 (3~4인분)	닭고기,오징어,밀,대두,닭고기,오징어,우유,대두,밀	['정제수']
3	국/반찬/메인 요리	누테이븐	메추리알 찜조림 350g	알류,메추리알,대두,밀,조개류,굴,알류,계란,우유,맥밀,땅콩,고등어,게,새우,돼지고...	['천일염', '염조간장']
4	국/반찬/메인 요리	모두의맛집	소문난 원조 조방낙지 낙곱새	대두,밀,새우,쇠고기,조개류,굴,난류,우유,메밀,땅콩,고등어,게,돼지고기,복숭아,토...	['마늘', '조개류', '전복', '대두유', '콩기름', '술알 포함']
...
3751	간식/과자/떡	청우	이슬 뽕 구워만든 죽염캔디 38g	땅콩,대두,우유,복숭아	['무엇', '자일리톨', '박하유']
3752	간식/과자/떡	키도컴	스테이버초콜렛	우유,오두,밀,메밀,대두	['...']
3753	간식/과자/떡	타르두프랑게	트뤼플렛 엑스티드 스위트 초콜렛 70g (떡)	밀,우유,대두,보리,땅콩,계란	['아몬드', '크린치 캐러멜', '준 초콜렛', '포도당 시럽', '헤이즐, ...']
3754	간식/과자/떡	kurly	맛다시마 젤리 100g (냉장)	없음	['설탕', '소금', '대두', '매실과육', '다시마', '간장']
3755	간식/과자/떡	마지언니네 빵맛간	밤고구마 시나몬 롤번	계란,우유,견과류,대두,밀	['설탕', '우유', '맥류', '견과류', '대두', '관색 설탕', '천일염, ...']

마켓컬리 데이터 전처리 완료 데이터 프레임

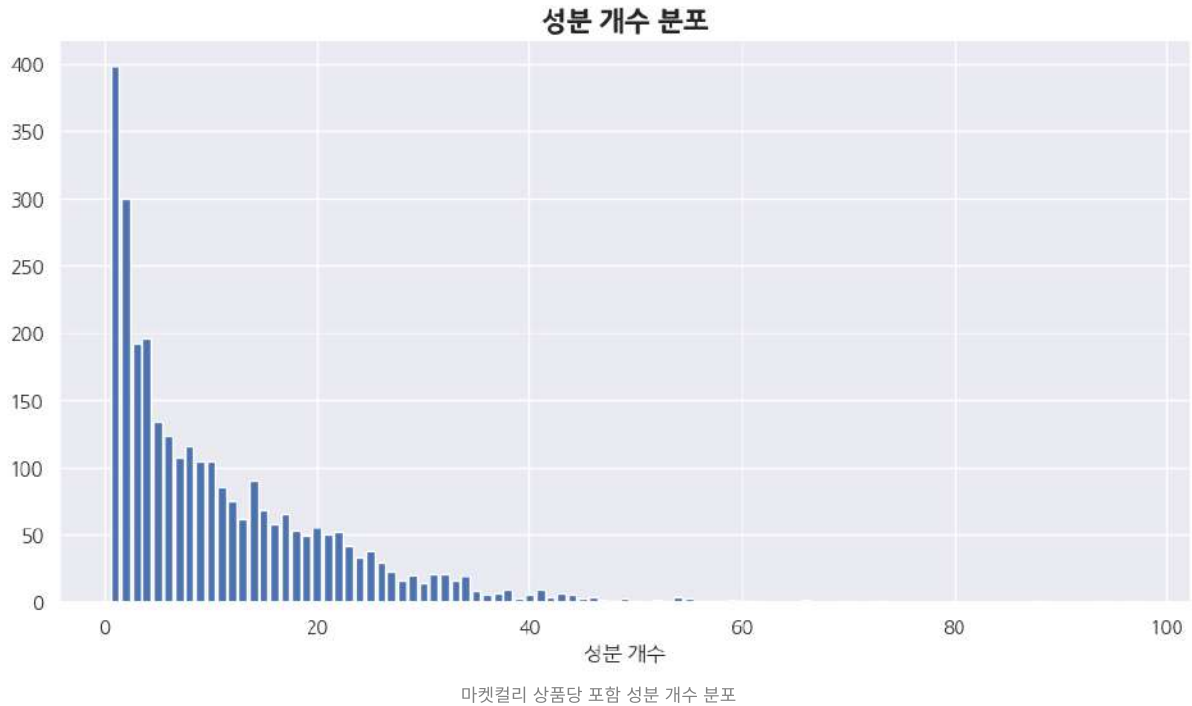
규칙을 찾아 성분들을 추출하기도 하고, 리서치와 텍스트 추출 과정을 통해 쌓아온 리스트를 활용해 성분들을 추출하는 과정을 거쳤지만 성분들을 추출하지 못한 데이터들이 존재하였다. 하지만 우리는 성분을 기준으로 데이터를 분류하고 추천할 예정이기 때문에 성분을 추출해 내지 못한 데이터는 삭제해 주었다.

이렇게 모든 분류와 추천 시스템에 활용할 2930개의 마켓컬리 데이터를 csv파일로 저장하였다.

	category	brand	item	allergy	ingredient
0	국/반찬/메인 요리	동우원	국산콩 나트 3종	국산콩생나트,대두,밀,쇠고기,검은콩생나트,대두,밀,쇠고기,국산콩와사비나트,대두,밀...	고등어,아황산류,맥밀,옥수수유옥수수,복숭아,오징어,술알포함,설탕,조개류,전복,대두...
1	국/반찬/메인 요리	kurly	전통 시장 물떡 (2팩)	밀,대두,우유,계란,우유,메밀,땅콩,고등어,게,새우,돼지고기,복숭아,토마토,아황산류...	조개류,정제소금,주정,맥류,조기,소스소스,참치,비타민,가공소금,오징어,구연산,복합조...
2	국/반찬/메인 요리	My little recipe	연등에서 올라온 전통찜닭 대용량 (3~4인분)	닭고기,오징어,밀,대두,닭고기,오징어,우유,대두,밀	정제수
3	국/반찬/메인 요리	누테이븐	메추리알 찜조림 350g	알류,메추리알,대두,밀,조개류,굴,알류,계란,우유,메밀,땅콩,고등어,게,새우,돼지고...	천일염,양조간장
4	국/반찬/메인 요리	모두의맛집	소문난 원조 조방낙지 낙곱새	대두,밀,새우,쇠고기,조개류,굴,난류,우유,메밀,땅콩,고등어,게,돼지고기,복숭아,토...	마늘,조개류,전복,대두유,콩기름,술알포함
...
3750	간식/과자/떡	이브투리에	분봉 초콜렛 12P	우유,대두,밀	아몬드,천연향료,유탄,천화당시럽,헤이즐,전분,코코아파우더,글리세린,프탈린,구연산,천...
3751	간식/과자/떡	청우	이슬 뽕 구워만든 죽염캔디 38g	땅콩,대두,우유,복숭아	무엇,자일리톨,박하유
3753	간식/과자/떡	타르두프랑게	트뤼플렛 엑스티드 스위트 초콜렛 70g (떡)	밀,우유,대두,보리,땅콩,계란	아몬드,크린치캐러멜,준초콜렛,포도당시럽,헤이즐,와이드초콜릿,코코아테스,유지...
3754	간식/과자/떡	kurly	맛다시마 젤리 100g (냉장)	없음	설탕,소금,대두,매실과육,다시마,간장
3755	간식/과자/떡	마지언니네 빵맛간	밤고구마 시나몬 롤번	계란,우유,견과류,대두,밀	설탕,우유,맥류,견과류,대두,관색설탕,천일염,계핏가루,정제수

마켓컬리 데이터 최종 데이터 프레임

한 상품당 얼마나 많은 성분들을 포함하고 있는지 살펴보기 위해 상품당 포함 성분 개수 분포를 그려보았다. 대부분 한 상품당 20개 미만의 성분들을 포함하고 있는 것을 확인할 수 있다.



iv. 분류 시스템 구현

마켓컬리 데이터를 비건단계에 따라 Flexitarian, Pesco-vegetarian, Lacto-ovo vegetarian, Lacto vegetarian, Ovo vegetarian, Vegan 총 6가지로 분류하는 시스템을 구현하고자 한다.



채식의 단계

돼지고기, 닭고기와 같은 육류가 하나라도 포함되어 있다면 Flexitarian 식품으로, 그 외 상품들 중 해산물이 포함되어 있다면 Pesco-vegetarian 식품으로, 그 외 식당들 중 우유와 동물 알류 포함되어 있다면 Lacto-ovo vegetarian 식품, 우유만 포함되어 있다면 Lacto vegetarian 식품, 동물 알류만 포함되어 있다면 Ovo vegetarian 식품, 모두 포함되지 않다면 Vegan 식품으로 분류할 것이다.

이 작업을 하기 위해서는 성분 리스트를 재료를 기준으로 나눈 개별 리스트들이 필요하다. 육류 성분 리스트(meat.txt), 해산물 리스트(sea.txt), 유제품 리스트(milk.txt), 알류 리스트(egg.txt)를 하나씩 불러와 리스트로 저장하였다.

```
# 분류 기준 성분 리스트 예시 - 육류
with open("./list/meat.txt", "r") as file:
    lst = file.readlines()

for word in lst:
    word=word.replace('\n','')
    if len(word)>0:
        kurly_meat.append(word)

meat_lst=[]
for word in kurly_meat:
    word=word.replace(' ','')
    meat_lst.append(word)

meat_lst=list(set(meat_lst))
```

먼저 가장 많은 종류의 재료를 섭취 가능한 Flexitarian 식품과 해산물까지 섭취 가능한 Pesco-vegetarian 식품을 먼저 분류해 주었다.

성분들 중 하나라도 육류 리스트에 포함되어 있다면 general으로 표시해 주었고, 그 외의 식품들 중 하나라도 해산물 리스트에 포함되어 있다면 pesco라고 표시해 주었다.

```
def classify_1(ingredient):
    # 일반식
    for word in ingredient.split(','):
        if word in meat_lst:
            return 'general'
    # 페스코
    for word in ingredient.split(','):
        if word in sea_lst:
            return 'pesco'

df['classify']=df['ingredient'].apply(classify_1)
```

Flexitarian 식품과 Pesco-vegetarian 식품으로 분류되지 않은 상품들을 따로 모아 Lacto-ovo vegetarian 식품, Lacto vegitarian 식품, Ovo vegetarian 식품으로 분류해주었다.

동물의 알류는 섭취하지 않고 유제품은 섭취하는 비건 단계와, 유제품은 섭취하지 않고 동물의 알류는 섭취하는 비건 단계, 두 개 모두 섭취하는 비건 단계로 나뉘지기 때문에 우선 유제품과 동물의 알류의 포함 여부 먼저 파악해야 했다.

```
def yes_milk(ingredient):
    for word in ingredient.split(','):
        if word in milk_lst:
            return '1'
    return '0'

def yes_egg(ingredient):
    for word in ingredient.split(','):
        if word in egg_lst:
            return '1'
    return '0'

df_need_classify['milk']=df_need_classify['ingredient'].apply(yes_milk)
df_need_classify['egg']=df_need_classify['ingredient'].apply(yes_egg)
```

이후 섭취 가능 식품에 맞게 Lacto-ovo, Lacto, Ovo vegetarian으로 분류해 주었다.

```
# 락토오보
df_need_classify.loc[(df_need_classify['milk']=='1')&(df_need_classify['egg']=='1'),'classify']='lacto_ovo'

# 락토
df_need_classify.loc[(df_need_classify['milk']=='1')&(df_need_classify['egg']=='0'),'classify']='lacto'

# 오보
df_need_classify.loc[(df_need_classify['milk']=='0')&(df_need_classify['egg']=='1'),'classify']='ovo'
```

이후 아무 곳에도 분류되지 않은 코드는 육류도, 해산물도, 유제품이나 동물의 알류 모두 포함하지 않은 식품이기 때문에 Vegan으로 분류해 주고 분류된 모든 데이터들을 합쳐 최종 데이터 프레임을 완성하였다.

```
df_need_classify['classify']='vegan'
df_final=pd.concat([df_classify,df_need_classify])
df_final.reset_index(drop=True,inplace=True)
```

	category	brand	item	allergy	ingredient	classify
0	국/반찬/메인 요리	풀무원	국산콩 나뭇잎 3종	국산콩,나뭇잎,대두,밀,쇠고기,검은콩,나뭇잎,대두,밀,쇠고기,국산콩,와사비,나뭇잎,대두,밀...	고등어,아황산류,메밀,옥수수,옥수수,복숭아,오징어,홍합포함,설탕,조개류,전분,대두...	general
1	국/반찬/메인 요리	kurly	전통 시장 물엿 (2팩)	밀,대두,우유,계란,우유,메밀,땅콩,고등어,게,새우,돼지고기,복숭아,토마토,아황산류...	조개류,정제소금,주정,액류,조기,소스,소스,참치,바타민,가공소금,오징어,구연산,복합...	pesco
2	국/반찬/메인 요리	모두의맛집	소문난 원조 조발낙지 낙금 새	대두,밀,새우,쇠고기,조개류,굴,난류,우유,메밀,땅콩,고등어,게,돼지고기,복숭아,토...	메밀,조개류,전분,내두유,콩기름,홍합포함	pesco
3	국/반찬/메인 요리	치즈음*테이스팅 룸	전복 술밥 리조토	전복,메밀,우유,조개류,전복,전복내장크림소스,대두,밀,우유,계란,닭고기,조개류,굴...	마늘,고등어,호두,유청,쇠고기,갯벌,포도당,구연산,가공아크릴,토코페롤,변성전분,주정...	general
4	국/반찬/메인 요리	피코크	금돼지식당 통상김장지찌개	대두,밀,돼지고기,쇠고기,알류,우유,메밀,땅콩,고등어,게,새우,복숭아,토마토,아황산...	메밀,고등어,황태중전제,굴산,호두,복숭아,홍합포함,연산이전분,닭고기,고춧가루,천연염...	general
...
2925	간식/과자/떡	라라스윗	발렌타인 기프트 세트	우유,밀,계란,대두,땅콩,호두,메밀,복숭아,토마토,돼지고기	알루미늄,당류	vegan
2926	간식/과자/떡	범스낵	그랜드 감자칩 2묵음 6종 (매1)	씨솔트,와사비,칠리앤드라이프,사우어크림앤드어니언,밀,우유,셀러리,겨자,사워크림앤드...	굴류,정제소금,오일,고추냉이,정제수정제소금,감자플레이크,리브뉴클레오타이드,나트륨,감자...	vegan
2927	간식/과자/떡	리터 스포트	메니 초콜릿믹스 9P	없음	초콜릿,옥수수	vegan
2928	간식/과자/떡	창우	야옹 뽀 구워만든 죽염앤디 38g	땅콩,대두,우유,복숭아	우유,자일리톨,백옥유	vegan
2929	간식/과자/떡	kurly	맛다시매 켄리 100g (냉장)	없음	설탕,소금,대두,매실과육,다시마,간장	vegan

마켓컬리 데이터 분류 완료 데이터프레임

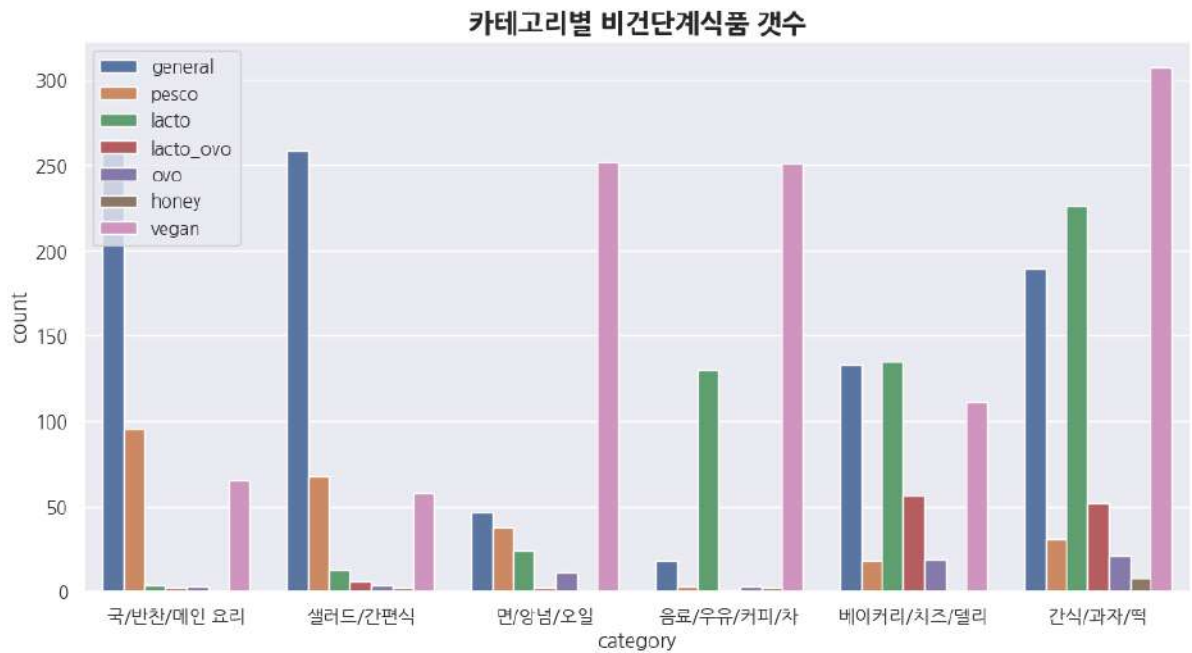


마켓컬리 분류 데이터 결과 시각화-비건단계식품 종류 및 개수 비교

그 결과 마켓컬리 데이터가 909개의 Flexitarian 식품, 253개의 Pesco vegetarian 식품, 532개의 Lacto vegetarian 식품, 119개의 Lacto-ovo vegetarian 식품, 61개의 Ovo vegetarian 식품, 1044개의 Vegan 식품으로 분류된 것을 확인할 수 있다.

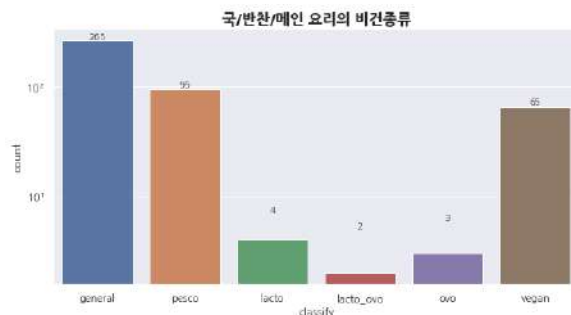
생각보다 Vegan에 해당하는 식품이 많았는데, 이는 이미지에서 텍스트로 추출하는 과정에서 모든 성분 데이터가 올바르게 글자를 인식하지 못하고 다른 단어로 인식되어서 발생한 결과라고 생각된다. 또한 분류하는 과정에서 정해진 기준에 해당하지 않은 남은 식품들이 Vegan으로 처리한 점도 Vegan으로 제일 많이 분류된 결과라고 생각된다. 이는 이미지에서 텍스트로 변환하는 과정에서 조금 더 정확도를 높이는 방법을 고안해낸다면 해결할 수 있는 문제라고 생각된다.

카테고리별로 비건단계의 식품들의 수를 비교해보았다.

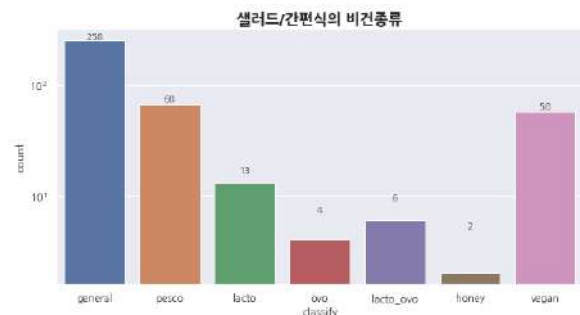


마켓컬리 분류 데이터 결과 시각화 - 카테고리별 비건 단계식품 개수 비교

한 카테고리씩 자세히 살펴보자. 먼저 국/반찬/메인 요리 카테고리 and 샐러드/간편식 비건 종류이다.



마켓컬리 분류 데이터 결과 시각화 - 국/반찬/메인요리



마켓컬리 분류 데이터 결과 시각화 - 샐러드/간편식

국/반찬/메인요리 카테고리 and 샐러드/간편식 카테고리 모두 Flexitarian식품의 상품이 제일 많았고 그 다음으로 Pesco, Vegan이었다.

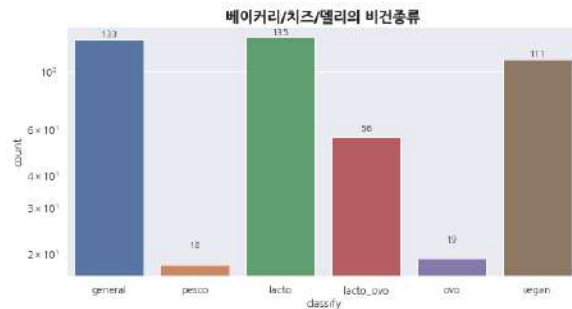


마켓컬리 분류 데이터 결과 시각화 - 면/양념/오일



마켓컬리 분류 데이터 결과 시각화 - 음료/우유/커피/차

면/양념/오일 카테고리 and 음료/우유/커피/차 카테고리는 Vegan 식품이 제일 많은 양을 차지하고 있었다. 두 카테고리 특성상 여러 음식재료들이 함께 있기보다는 한 성분 자체로 높은 비중을 차지하는 식품들이 많기 때문에 나타난 결과라고 판단된다. 또한 음료/우유/커피/차 카테고리에서는 Lacto vegetarian의 비중이 높았는데, 유제품인 우유가 포함되어 있어 나타난 결과라고 판단된다.



마켓컬리 분류 데이터 결과 시각화 - 베이커리/치즈/델리



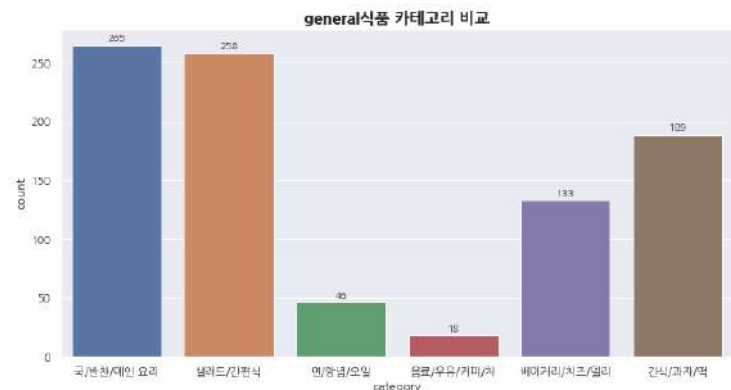
마켓컬리 분류 데이터 결과 시각화 - 간식/과자/떡

베이커리/치즈/델리 카테고리 and 간식/과자/떡 카테고리의 식품들에서는 다른 카테고리보다 Lacto vegeterain 식품과 Lacto-ovo vegetarian 식품이 높은 비중을 차지하고 있었다. 이는 유제품이 많이 사용되는 제품이다 보니 나타난 결과라고 보인다. 또한 간식/과자/떡에서 많은 상품이 Vegan 식품으로 분류되었는데, 곡물을 주재료로 만든 간식들이 많이 있어 나타난 결과로 보인다.

이번에는 분류 결과를 비건 단계 별로 살펴보자. 먼저 Flexitarian식품이다.



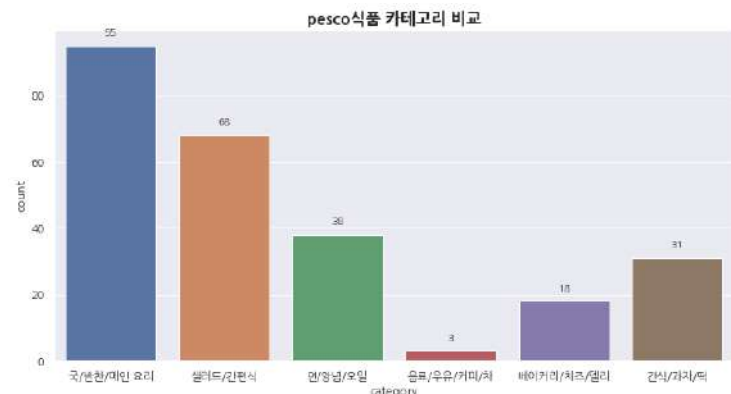
마켓컬리 분류 데이터 결과 시각화 - flexiterian 워드클라우드



마켓컬리 분류 데이터 결과 시각화 - flexiterian 카테고리 비교

Flexitarian 식품으로 분류된 상품을 중 육류 리스트에 포함되는 단어들을 고기 모양으로 워드 클라우드도 만들어 보았다. 많은 상품들이 국/반찬/메인요리와 샐러드/간편식으로 분류된 것을 확인할 수 있다. 메인 요리로 많이 사용되는 닭고기, 돼지고기, 쇠고기 등과 함께 샐러드에 재료로 많이 사용되는 닭가슴살 등이 눈에 띄는 것을 확인할 수 있고, 베이커리나 간식류를 만들 때 사용되는 젤라틴도 포함되어 있는 것을 확인할 수 있다.

Pesco vegetarian 식품으로 분류된 상품들의 카테고리를 살펴보자.

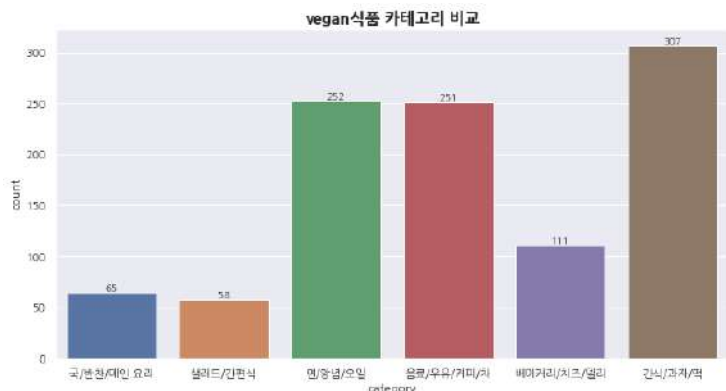


마켓컬리 분류 데이터 결과 시각화 - pescos 카테고리 비교



Ovo 식품으로 분류된 상품들 중 알류 리스트에 포함되는 단어들로 알 모양 워드 클라우드로 함께 만들었다. 다른 비건 단계 식품을 보다 분류된 수는 적지만 베이커리/치즈/델리 카테고리 간식/과자/떡 카테고리에서는 다른 카테고리보다 비교적 많은 상품을 가지고 있는 것을 확인할 수 있다.

마지막으로 Vegan식품으로 분류된 상품들의 카테고리이다.



Vegan 식품으로 분류된 상품들의 성분들을 사용해 새싹 모양 워드 클라우드도 함께 만들었다. 대두, 사과, 땅콩, 옥수수과 같은 과일, 채소 등도 눈에 띄지만, 당류, 정제소금 등과 같은 재료들이 눈에 띈다. 앞에서 다른 비건 단계 식품으로 분류되지 않은 상품들의 성분 정보이다 보니 주 재료들보다 다른 성분들이 많은 비중을 차지하고 있는 것으로 파악된다. 밀가루, 쌀과 같은 곡물들이 주로 사용되는 간식/과자/떡에 많은 상품들이 있으며, 주 재료 하나로 맛을 내는 면/양념/오일과 음료/우유/커피/차 카테고리에도 많은 상품들이 있는 것을 확인할 수 있다.

v. 추천 시스템 구현

전처리와 분류를 마친 마켓컬리 데이터를 불러와서 추천 시스템을 구현해 보았다.

	category	brand	item	allergy	ingredient	classify
0	국/반찬/케일 요리	물푸름	국산콩 니루 3종	국산콩산나트,태두,분쇄고기,검은콩산나트,태두,분쇄고기,국산콩와사비나트,태두,분쇄...	고동여,아황산류,메밀,옥수수유옥수수,복숭아,오징어,홍합포함,설탕,조개류,천료...	general
1	국/반찬/케일 요리	kurly	전통 시장 물먹 (2매)	밀,태두,우유,케란,우유,메밀,땅콩,고동여,게,새우,돼지고기,복숭아,도마토,야...	조개류,정제소금,우정,떡류,조기,소소스,참치,바터런,가공소금,오징어,구운산,죽...	pesco
2	국/반찬/케일 요리	모두의맛집	소문난 원조 조항나지 낙골	태두,밀,새우,쇠고기,조개류,굴,나무,우유,메밀,땅콩,고동여,게,돼지고기,복숭아,...	매늘,조개류,전복,새우류,홍기름,홍합포함	pesco
3	국/반찬/케일 요리	자드름x테이스팅	전복 순살 리조토	전복마역밀,우유,조개류,전복,전복나탕크립스,태두,분,우유,케란,덜고기,조개...	매늘,고동여,초두,양친,소고기,젓갈,포도당,구운산,가공아크인,도코페놀,변성전...	general
4	국/반찬/케일 요리	피코리	금돼지색만 통산건전지채	태두,물,돼지고기,쇠고기,알류,우유,메밀,땅콩,고동여,게,새우,복숭아,도마토,야...	매늘,고동여,함대콩진제,굴,산나트,복숭아,홍합포함,인산이진제,달고미,고춧가루...	general
-	-	-	-	-	-	-
2525	간식/과자/떡	라리스엡	볼렌타인 기프트 세트	우유,분,케란,태두,땅콩,조개,메밀,복숭아,포도당,돼지고기	알몬로스,당류	vegan
2526	간식/과자/떡	벌스넥	그랜드 감자칩 2종류 6종 (메1)	씨솔트,와사비,칠리엔드라인,사우어크림엔드어니인,물,우유,살리리,겨자,사리...	글루,정제소금,오일,고추냉이,정제수정제소금,감자플레이크,리브,부농글루오터디...	vegan
2527	간식/과자/떡	리더 스포트	메니 초콜렛믹스 9g	없음	나트륨,감자...	vegan
2528	간식/과자/떡	청우	아름 변 구워민든 죽염캔디 38g	없음	훈초밀엡,옥수수	vegan
2529	간식/과자/떡	kurly	맛다시마 분리 100g (냉장)	없음	무엡,자일리톨,박하엡	vegan
2530	간식/과자/떡				설탕,소금,태두,매실과육,다시마,간장	vegan
2930 rows x 6 columns						

상품을 추천하는 기준은 식품의 성분이다. 따라서 먼저 CountVectorizer를 사용하여 성분들을 벡터화 시켜주었다.

```
from sklearn.feature_extraction.text import CountVectorizer

df['ingredient']=df['ingredient'].apply(lambda x: x.replace(',',' '))
count_vect=CountVectorizer(min_df=0, ngram_range=(1,2))
ingredient_mat=count_vect.fit_transform(df['ingredient'])
```

이후 코사인 유사도를 사용하여 품 성분의 텍스트 유사도를 파악해 주고 유사도가 높은 순으로 정렬해 주었다.

```
from sklearn.metrics.pairwise import cosine_similarity

ingredient_sim=cosine_similarity(ingredient_mat,ingredient_mat)
ingredient_sim_sorted_ind=ingredient_sim.argsort()[::-1]
```

이후 사용자가 선택할 수 있는 다양한 옵션을 주어 상황에 맞는 상품을 추천해 주는 함수를 만들었다.

먼저 하나는 카테고리 옵션으로, 사용자가 카테고리 옵션을 선택한다면 입력한 식품과 같은 카테고리의 상품만을 추천해 주고, 카테고리 옵션을 선택하지 않는다면 카테고리에 상관없이 비슷한 성분을 가진 식품을 추천해 준다.

다른 하나는 비건 옵션이다. 비건 옵션으로 나올 수 있는 결과는 가지이다.

1. 1번 비건 옵션 : 성분이 비슷한 상품들 중 같은 비건 단계의 상품들만 추천해 준다.
2. 2번 비건 옵션 : 성분이 비슷한 상품들을 각 비건단계마다 2개씩 추천해 준다.
3. 3번 비건 옵션 : 성분이 비슷한 상품들 중 Vegan 단계의 식품들만 추천해 준다.
4. 비건 옵션을 선택하지 않은 경우 : 비건 단계에 상관없이 성분이 비슷한 상품들을 추천해 준다.

물론 카테고리 옵션과 비건 옵션을 동시에 적용 가능하도록 하였다.

```
def find_sim_ingredient(df, sorted_ind, product_name, top_n=10, cate=0, vegan=0):

    title_product=df[df['item']==product_name]
    title_index=title_product.index.values

    similar_indexes=sorted_ind[title_index]
    similar_indexes=similar_indexes[similar_indexes!=title_index]

    similar_indexes=similar_indexes.reshape(-1)
    option_lst=similar_indexes

    # 카테고리 옵션
    if cate==1:
        option_lst=[]
        category=df.loc[df['item']==product_name, 'category'].values
        cate_indexes=df.loc[df['category']==category[0]].index.values

    # 비건단계 옵션1 - 같은 종류의 단계만 추천
    if vegan==1:
        option_lst=[]
```

```

vegan_level=df.loc[df['item']==product_name,'classify'].values
same_vegan_indexes=df.loc[df['classify']==vegan_level[0]].index.values

# 비건단계 옵션2 -비건단계당 2개씩 추천
elif vegan==2:
    option_lst=[]
    vegan_check={'pesco':0,'lacto_ovo':0,'lacto':0,'ovo':0,'vegan':0,'honey':0}
    cnt=0

# 비건단계 옵션3 - 비건 제품만 추천
elif vegan==3:
    option_lst=[]
    vegan_indexes=df.loc[df['classify']=='vegan'].index.values

# 옵션 적용
if cate==1 and vegan==0: # 카테고리
    for idx in similar_indexes:
        if idx in cate_indexes:
            option_lst.append(idx)

elif cate==0 and vegan==1: # 같은비건단계
    for idx in similar_indexes:
        if idx in same_vegan_indexes:
            option_lst.append(idx)

elif cate==0 and vegan==2: # 비건단계별
    for idx in similar_indexes:
        vegan_level=df.iloc[idx,-1]
        if vegan_level!='general' and vegan_check[vegan_level]<2:
            vegan_check[vegan_level]+=1
            option_lst.append(idx)
            cnt+=1
        elif cnt==12:
            break
        else:
            continue

elif cate==0 and vegan==3: # 비건제품만
    for idx in similar_indexes:
        if idx in vegan_indexes:
            option_lst.append(idx)

elif cate==1 and vegan==1: # 카테고리, 같은비건단계
    for idx in similar_indexes:
        if idx in cate_indexes and idx in same_vegan_indexes:
            option_lst.append(idx)

elif cate==1 and vegan==2: # 카테고리, 비건단계별
    for idx in similar_indexes:
        if idx in cate_indexes:
            vegan_level=df.iloc[idx,-1]
            if vegan_level!='general' and vegan_check[vegan_level]<2:
                vegan_check[vegan_level]+=1
                option_lst.append(idx)
                cnt+=1
            elif cnt==12:
                break
            else:
                continue

elif cate==1 and vegan==3: # 카테고리, 비건제품만
    for idx in similar_indexes:
        if idx in cate_indexes and idx in vegan_indexes:
            option_lst.append(idx)

return df.iloc[option_lst][:top_n].sort_values('classify')

```

작동 예시를 살펴보면 다음과 같다.

비비고의 총각김치를 추천시스템에 입력해보겠다.

	category	brand	item	allergy	ingredient	classify
10	국/반찬/메인 요리	비비고	총각김치1.5KG	새우,대두,밀,조개류,굴,전복,홍합포함,젓	마늘 조개류 식염 고추 절임알타리우 배류레 감칠맛엑젓 고춧가루 하선정남해인명명멸시엑...	pesco

마켓컬리 추천시스템 예시 입력 상품

아무 옵션도 선택하지 않고 추천시스템을 돌린 결과는 다음과 같다.

	category	brand	item	allergy	ingredient	classify
172	국/반찬/메인 요리	비비고	종각 김치	새우,대두	마늘 조개류 식염 고추 대두함유 절임알타리무 배류레 감칠맛엑젱 고춧가루 멸치 김치 총각무	pesco
303	국/반찬/메인 요리	비비고	포기 배추김치 4.9kg	밀,조개류,굴,전복,홍합포함,젓,새우,대두	절임배추 배추 이기지간마늘 마늘 조개류 식염 고추 쪽파 고춧가루 배류레 채소 새우액...	pesco
107	국/반찬/메인 요리	비비고	별은 배추김치	새우,대두	절임배추 배추 유산균배양액배류레 식염 고추 새우액젱 채소류 감칠맛엑젱 고춧가루 멸치...	pesco
278	국/반찬/메인 요리	비비고	별은배추김치3kg	새우,대두,밀,조개류,굴,전복,홍합포함,젓	절임배추 배추 마늘 식염 고추 쪽파 채소류 대파 대두함유 멸치 생강 새우액젱 김치 ...	pesco
63	국/반찬/메인 요리	비비고	파김치 400g	밀,조개류,굴,전복,홍합포함,젓,새우,대두	양배 마늘 생강 이별치조미엑젱 멸치엑젱 식염 전단건 고추 조개류 전복 고춧가루 감...	pesco
217	국/반찬/메인 요리	비비고	석박지 900g	새우,대두,밀,조개류,굴,전복,홍합포함,젓성분성있음	마늘 간마늘 소르비톨 멸치 멸치엑젱 감칠맛엑젱 기타가공품 생강 결절과당 젖산칼슘...	pesco
724	면/양념/오일	선물세트	파르키오니 포도씨유	없음	식당 식염 고추 대파 기타가공품 고춧가루 멸치 물엿 올고추	pesco
31	국/반찬/메인 요리	비비고	별은 배추김치 1.8kg	새우,대두,밀,조개류,굴,전복,홍합포함,젓	마늘 대두함유 소르비톨 멸치 절임배추 배추 양배 쪽파 배류레 김치 당류 생강 김치양...	general
2074	면/양념/오일	비비도기찬	저칼로리 비명장	우유,대두,밀,쇠고기,달걀,매밀,참깨,고등어,게,새우,돼지고기,복숭아,마마도,아황산...	고춧가루 마늘 주정 양파	vegan
93	국/반찬/메인 요리	비비고	별은배추김치 더중후한맛 900g	새우,대두,밀,조개류,굴,전복,홍합포함,젓	마늘 변태이젓변태이 간마늘 배류레새우 대두함유 멸치 절임배추 배추 조개류 김치속...	general

아무 옵션도 선택하지 않은 경우

주로 입력한 상품과 비슷한 종류인 김치 종류들을 많이 추천해 주었고, 그렇지 않은 상품에 대해서는 입력한 상품의 성분들과 비슷한 성분들로 구성된 식품을 추천해 주는 것을 확인할 수 있다.

카테고리 옵션만 선택한 결과는 다음과 같다.

	category	brand	item	allergy	ingredient	classify
172	국/반찬/메인 요리	비비고	종각 김치	새우,대두	마늘 조개류 식염 고추 대두함유 절임알타리무 배류레 감칠맛엑젱 고춧가루 멸치 김치 총각무	pesco
303	국/반찬/메인 요리	비비고	포기 배추김치 4.9kg	밀,조개류,굴,전복,홍합포함,젓,새우,대두	절임배추 배추 이기지간마늘 마늘 조개류 식염 고추 쪽파 고춧가루 배류레 채소 새우액...	pesco
107	국/반찬/메인 요리	비비고	별은 배추김치	새우,대두	절임배추 배추 유산균배양액배류레 식염 고추 새우액젱 채소류 감칠맛엑젱 고춧가루 멸...	pesco
278	국/반찬/메인 요리	비비고	별은배추김치3kg	새우,대두,밀,조개류,굴,전복,홍합포함,젓	절임배추 배추 마늘 식염 고추 쪽파 채소류 대파 대두함유 멸치 생강 새우액젱 김치 ...	pesco
63	국/반찬/메인 요리	비비고	파김치 400g	밀,조개류,굴,전복,홍합포함,젓,새우,대두	양배 마늘 생강 이별치조미엑젱 멸치엑젱 식염 전단건 고추 조개류 전복 고춧가루 감...	pesco
217	국/반찬/메인 요리	비비고	석박지 900g	새우,대두,밀,조개류,굴,전복,홍합포함,젓성분성있음	마늘 간마늘 소르비톨 멸치 멸치엑젱 감칠맛엑젱 기타가공품 생강 결절과당 젖산칼슘...	pesco
31	국/반찬/메인 요리	비비고	별은 배추김치 1.8kg	새우,대두,밀,조개류,굴,전복,홍합포함,젓	마늘 대두함유 소르비톨 멸치 절임배추 배추 양배 쪽파 배류레 김치 당류 생강 김치양...	general
93	국/반찬/메인 요리	비비고	별은배추김치 더중후한맛 900g	새우,대두,밀,조개류,굴,전복,홍합포함,젓	마늘 변태이젓변태이 간마늘 배류레새우 대두함유 멸치 절임배추 배추 조개류 김치속...	general
216	국/반찬/메인 요리	비비고	멸우김치	새우,대두	마늘 유산균배양액배류레 식염 고춧 고춧파 대파 멸치통시름 대두함유 소르비톨 정제수...	general
2	국/반찬/메인 요리	모두의맛 집	소문난 원조 조깅낙지 낙곱새	대두,밀,새우,쇠고기,조개류,굴,난류,우유,매밀,참깨,고등어,게,돼지고기,복숭아,마...	마늘 조개류 전복 내두유 콩기름 홍합포함	pesco

카테고리 옵션 선택

카테고리 옵션을 선택하지 않았을 경우에는 면/양념/오일 카테고리에서도 상품이 추천된 것과 달리 카테고리 옵션을 선택한 경우 국/반찬/메인요리 카테고리에서만 상품을 추천해 주는 것을 확인할 수 있다.

같은 비건 단계의 상품만 추천해주는 1번 비건 옵션을 선택한 결과이다.

	category	brand	item	allergy	ingredient	classify
172	국/반찬/메인 요리	비비고	종각 김치	새우,대두	마늘 조개류 식염 고추 대두함유 절임알타리무 배류레 감칠맛엑젱 고춧가루 멸치 김치 총각무	pesco
303	국/반찬/메인 요리	비비고	포기 배추김치 4.9kg	밀,조개류,굴,전복,홍합포함,젓,새우,대두	절임배추 배추 이기지간마늘 마늘 조개류 식염 고추 쪽파 고춧가루 배류레 채소 새우액...	pesco
107	국/반찬/메인 요리	비비고	별은 배추김치	새우,대두	절임배추 배추 유산균배양액배류레 식염 고추 새우액젱 채소류 감칠맛엑젱 고춧가루 멸...	pesco
278	국/반찬/메인 요리	비비고	별은배추김치3kg	새우,대두,밀,조개류,굴,전복,홍합포함,젓	절임배추 배추 마늘 식염 고추 쪽파 채소류 대파 대두함유 멸치 생강 새우액젱 김치 ...	pesco
63	국/반찬/메인 요리	비비고	파김치 400g	밀,조개류,굴,전복,홍합포함,젓,새우,대두	양배 마늘 생강 이별치조미엑젱 멸치엑젱 식염 전단건 고추 조개류 전복 고춧가루 감...	pesco
217	국/반찬/메인 요리	비비고	석박지 900g	새우,대두,밀,조개류,굴,전복,홍합포함,젓성분성있음	마늘 간마늘 소르비톨 멸치 멸치엑젱 감칠맛엑젱 기타가공품 생강 결절과당 젖산칼슘...	pesco
724	면/양념/오일	선물세트	파르키오니 포도씨유	없음	식당 식염 고추 대파 기타가공품 고춧가루 멸치 물엿 올고추	pesco
2	국/반찬/메인 요리	모두의맛 집	소문난 원조 조깅낙지 낙곱새	대두,밀,새우,쇠고기,조개류,굴,난류,우유,매밀,참깨,고등어,게,돼지고기,복숭아,마...	마늘 조개류 전복 내두유 콩기름 홍합포함	pesco
324	국/반찬/메인 요리	진가네반찬	다시마완	대두,밀,고등어,게,새우,돼지고기,호두,닭고기,쇠고기,오징어,조개류	천연향신료 식염 천양고추 고춧가루 멸치 올고추	pesco
680	면/양념/오일	Kurly's	요리용 맑은 육수 1L	알룰,달걀,우유,참깨,대두,밀,쇠고기,조개류,배지락,기초계,홍합포함,	마늘 고추 배지락 멸치 천양고추 키초계 건표고버섯,당류	pesco

1번 비건 옵션 선택

비건 옵션을 선택하지 않았을 때는 pesco와 general로 분류된 식품들이 추천되었지만, 1번 비건 옵션을 선택한 경우에는 입력한 상품의 비건 단계의 pesco와 같은 비건 단계 식품들만을 추천해 주는 것을 확인할 수 있다.

이번에는 각 비건 단계당 2개씩 추천해주는 2번 비건 옵션을 선택한 결과이다.

	category	brand	item	allergy	ingredient	classify
172	국/반찬/메인 요리	베비고	종각 김치	새우,대두	마늘 조개류 식염 고추 대두쌀유 절임알타리무 배류래 감실맛액젓 고춧가루 멸치 김치 종각무	pesco
303	국/반찬/메인 요리	베비고	포기 배추김치 4.9kg	밀,조개류,굴,전복,홍합포함,젓,새우,대두	절임배추 배추 이기지킨마늘 마늘 조개류 식염 고추 쪽파 고춧가루 배류래 채 소 새우액...	pesco
2074	면/양념/오일	비비드키친	지칼로리 비빔장	우유,대두,밀,쇠고기,달걀,맥분,땅콩,고등어,게,새우,돼지고기,복숭 아,토마토,아황산...	고춧가루 마늘 주황 양파	vegan
2028	면/양념/오일	햇님마을	국산 100% 청양 고춧가루 2중 (22년산 햇 고춧가루)	없음	고춧가루	vegan
1327	베이커리/치즈/밀 리	서울우유	슈레드 모짜렐라 300g	우유	우유용고효소 자연지즈 식염 포도당	lacto
1120	국/반찬/메인 요리	스마일찬	전미재 간장볶음	대두,밀,오징어,게란,쇠고기,고등어,게,새우,돼지고기,호두,닭고기,조 개류	마늘 조미건어포류 알지태도 게란	ovo
1416	베이커리/치즈/밀 리	상학치즈	브라 자연치즈	우유,대두,땅콩,밀,게란	당류 식염 찹은 유산균	lacto
1143	샐러드/간편식	오뚜기	참깨리얼 (115GX4)	밀,대두,게란,우유,쇠고기,돼지고기,닭고기,오징어,조개류,굴,홍합포 함...	마늘 한창깨게란블록 천분 땅이 건파 감실맛베이스 구아검 쇠고기육수분말 소맥분 마늘치...	ovo
1137	샐러드/간편식	고매공방	쁘띠봉어빵 2중	밀,대두,게란,우유,밀유,밀,대두,게란,우유,메밀,땅콩,고등어,게,새우, 돼지고기,복...	우유 설탕 식염 마가린 대두 기타가공품 메밀 알가루 땅이 게란 타피오카 찹 수수	lacto_ovo
1352	베이커리/치즈/밀 리	설탕없는과자 공방	가벼운 산소빵 4중(5개입)	산소빵,대두,밀,게란,우유,땅콩,돼지고기,토마토,호두,닭고기,여니언 크림,대두,우유...	호두 자른 난소화성알토덱스트린 알룰로스 가공이크림 소트비탄지방산소 유산균주 합치로...	lacto_ovo

2번 비건 옵션 선택

입력한 단계인 pesco뿐만 아니라 general, lacto-ovo, lacto, ovo, vegan 단계의 식품들을 2개씩 추천해 주고 있는 것을 확인할 수 있다.

마지막으로 Vegan 식품만 추천해주는 3번 옵션을 선택한 결과이다.

	category	brand	item	allergy	ingredient	classify
2074	면/양념/오일	비비드키친	지칼로리 비빔장	우유,대두,밀,쇠고기,달걀,맥분,땅콩,고등어,게,새우,돼지고기,복숭아,토마 토,아황산...	고춧가루 마늘 주황 양파	vegan
2028	면/양념/오일	햇님마을	국산 100% 청양 고춧가루 2중 (22년산 햇 고춧가루)	없음	고춧가루	vegan
1925	국/반찬/메인 요리	조선호밀김 치	갈치식백지	새우,우유,대두,쇠고기,조개류,전복,밀,젓	마늘 식염 쪽파 고추 청제수 생강 아스타틴 김매로 멸무	vegan
1935	국/반찬/메인 요리	조선호밀김 치	조선 주니어 배추김치 400g	새우,우유,대두,쇠고기,조개류,전복,밀,젓	마늘 상갈 식염 사과 산불가루 고춧가루 부추 집합물 새우젓무 설탕올리 고당 양파	vegan
2668	간식/과자/떡	미식이 부각	김부각	없음	마늘 옥수수 천분 기타가공품 유착유 양파	vegan
2081	면/양념/오일	햇님마을	명양산 고춧가루 1kg (22년도 햇 고춧가루)	없음	고춧가루 건고추	vegan
1940	국/반찬/메인 요리	조선호밀김 치	조선 주니어 백작두기 400g	새우,우유,대두,쇠고기,조개류,전복,밀,젓	마늘 버드	vegan
2065	면/양념/오일	산올리아노	페페로치노 25g	없음	건고추 고추	vegan
2148	면/양념/오일	해산물	물로만 끓여도 되는 된장짜개 3중	자른,대두,밀,쇠고기,청양조,대두,밀,조개류,배지락,배지락맛,대두,밀,새 우,조개류...	당류 고추	vegan
2619	베이커리/치즈/밀 리	멜로조	아메리코악장 슬라이스 60g	돼지고기	식염 보존료	vegan

3번 비건 옵션 선택

Vegan으로 분류된 식품들만을 추천해주고 있는 것을 확인할 수 있다.

카테고리 옵션과 비건 옵션 중 하나인 2번 옵션을 선택한 결과는 다음과 같다.

	category	brand	item	allergy	ingredient	classify
172	국/반찬/메인 요리	베비고	종각 김치	새우,대두	마늘 조개류 식염 고추 대두쌀유 절임알타리무 배류래 감실맛액젓 고춧가루 멸치 종각무	pesco
303	국/반찬/메인 요리	베비고	포기 배추김치 4.9kg	밀,조개류,굴,전복,홍합포함,젓,새우,대두	절임배추 배추 이기지킨마늘 마늘 조개류 식염 고추 쪽파 고춧가루 배류래 채 소 새우액...	pesco
1925	국/반찬/메인 요리	조선호밀김치	갈치식백지	새우,우유,대두,쇠고기,조개류,전복,밀,젓	마늘 식염 쪽파 고추 청제수 생강 아스타틴 김매로 멸무	vegan
1935	국/반찬/메인 요리	조선호밀김치	조선 주니어 배추김치 400g	새우,우유,대두,쇠고기,조개류,전복,밀,젓	마늘 상갈 식염 사과 산불가루 고춧가루 부추 집합물 새우젓무 설탕올리 고당 양파	vegan
1120	국/반찬/메인 요리	스마일찬	전미재 간장볶음	대두,밀,오징어,게란,쇠고기,고등어,게,새우,돼지고기,호두,닭고기,조개 류	마늘 조미건어포류 알지태도 게란	ovo
1121	국/반찬/메인 요리	해터의부엌x더반 찬	제주불소라 미역국	우유,메밀,대두,밀,게,새우,조개류,소라,전복,홍합,알류,땅콩,고등어,돼지고기, 복숭...	대과 미역 토마토 복숭아 전복소스분향 유지방 양파	lacto
1115	국/반찬/메인 요리	비글반찬	미역줄기볶음	대두,닭고기,알류,우유,메밀,땅콩,밀,고등어,게,새우,돼지고기,복숭아,토마토, 아황산...	우유 대두유 정어 대과 다진 절임미역볶기 미역 메밀 글소수지 양파	lacto
1116	국/반찬/메인 요리	빔스	크로콘올라 피자	밀,대두,우유,쇠고기,게란,새우,땅콩,호두,닭고기,복숭아,오징어,고등어,게 릴,아황산...	빵류 피자치즈 마요네즈	lacto_ovo
1123	국/반찬/메인 요리	글루스 다이닝	브라조볼	달걀,우유,대두,밀,돼지고기,토마토,닭고기,쇠고기,조개류,배지락,모시조개, 메밀,땅콩...	버터원크 땅이 청제수 식용염유지 난향 난향계 기타가공품 알가루 현미식물 우유 소금 ...	lacto_ovo
1117	국/반찬/메인 요리	얼큰몰에	함출은 국산콩 순두부 2 개입	대두,우유,게란,밀,땅콩	곤합제제 구연산상나트륨 게란	ovo

카테고리 옵션 2번 비건 옵션 모두 선택

국/반찬/메인요리 카테고리 상품들 중 Pesco, General, Lacto-ovo, Lacto, Ovo, Vegan 상품을 각각 2개씩 추천해 주고 있는 것을 확인할 수 있다.

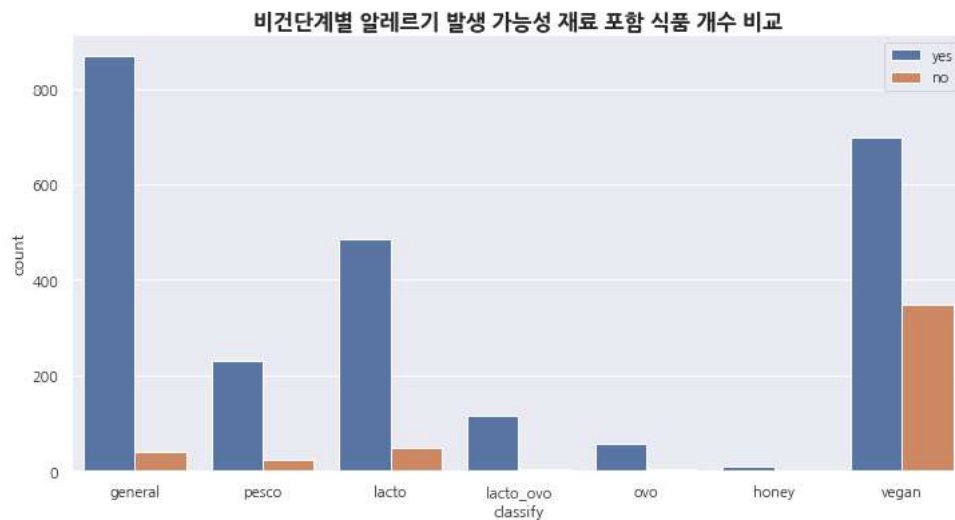
vi. 알레르기 및 글루텐 포함 여부 파악

알레르기를 가지고 있는 사람들은 음식을 섭취하기 전에 먹으려고 하는 음식에 알레르기가 포함되어 있는지 먼저 확인을 하고 섭취한다. 또한 일반적으로 알려진 알레르기뿐만 아니라 밀가루 알레르기도 등장하게 되면서 글루텐의 포함 여부도 파악하고자 하는 사람들이 등장하고 있다.

따라서 성분을 이용해서 분류하고 추천 시스템을 구현하는 과정에서 알레르기 성분과 글루텐 포함 여부를 함께 알려준다면 더 좋은 서비스가 될 것이라고 판단하여 이 기능을 추가하였다.

알레르기 유발 성분이 식품에 포함되어 있으면 알레르기 성분이 포함되어 있다는 메시지를 보여주기로 하였다. 하지만 식품에서 알레르기를 일으키는 종류는 워낙 다양하기도 하고 사람마다 모두 다르기 때문에 처음부터 보여주기로 하는 것보다는 사용자가 알레르기에 대한 정보를 입력받은 후에 안내 메시지를 출력하기로 하였다. 작동 예시는 다음 챗봇인 사용 예시에서 보여주도록 하겠다.

마켓컬리 데이터를 수집할 때 알레르기 정보를 따로 적어둔 페이지가 있어 크롤링 해왔었다. 이를 이용해 알레르기 발생 가능성이 있는 재료를 포함한 식품의 개수를 비건 단계별로 비교해 보는 그래프를 그려보았다.



마켓컬리 데이터 알레르기 시각화

가장 많은 종류의 성분 종류들을 섭취하는 Flexitarian vegetarian 식품이 알레르기 발생 가능성 재료가 가장 많이 포함하고 있는 것을 확인할 수 있었다.

반면 글루텐을 포함하고 있는 식품 및 성분은 '글루테닌', '글리아딘', '보리', '귀리', '밀', '밀가루', '종력분', '강력분', '박력분' 등으로 정해져 있다.

식품 성분들 중 여기에 해당하는 성분이 존재한다면 글루텐 column에 따로 추가하였고, 글루텐이 포함되지 않은 상품이라면 '발견사항 없음'이라고 표시해 주었다.

```
def gluten_ckeck(ingredient):
    gluten_lst=[]
    for word in ingredient.split(','):
        if word in glu:
            gluten_lst.append(word)
    if len(gluten_lst)>0:
        return ','.join(gluten_lst)
    else:
        return '발견사항 없음'

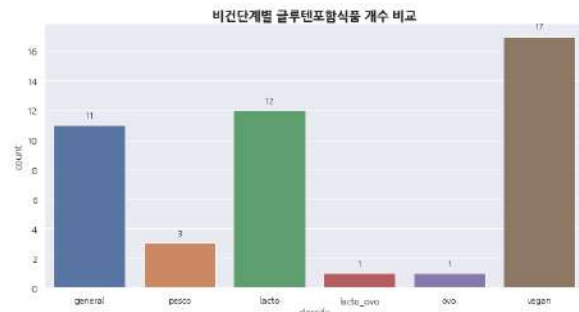
kurly['gluten']=kurly['ingredient'].apply(gluten_ckeck)
```

다음은 글루텐 포함 여부 파악까지 마친 마켓컬리 데이터 프레임이다.

[illegible]

마켓컬리 데이터 글루텐 포함 여부

비건 단계별로 글루텐 포함 여부를 비교해보았다.



마케팅컬리 데이터 글루텐 포함 식품 시각화



마케팅컬리 데이터 글루텐 미포함 식품 시각화

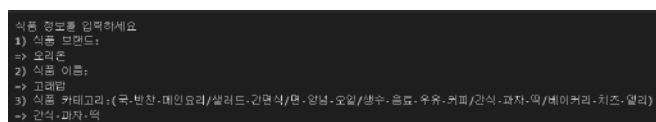
글루텐을 포함하고 있지 않은 식품들이 훨씬 많았지만 비건 단계별 비율은 비슷한 양상을 보이는 것을 확인할 수 있다.

vii. 사용 예시

앞에서 만들어낸 코드들을 활용해서 사용자가 새로운 식품을 입력했을 때 어떻게 진행되는지 그 흐름을 정리해 보았다.

1. 식품 정보 및 사진 입력

먼저 사용자에게 새로운 상품에 대한 정보를 입력 받는다.



마켓컬리 사용예시 상품정보입력



마케팅컬리 사용예시 상품정보 입력 gif

식품 성분은 사진을 찍어 입력받도록 한다. pytesseract를 통해 입력받은 이미지에서 텍스트를 추출하고 전처리를 거쳐 성분에 대한 데이터만 싹터준다.

[고래밥]에는 글루텐이 포함되어 있습니다.

마켓컬리 사용예시 글루텐 정보 결과

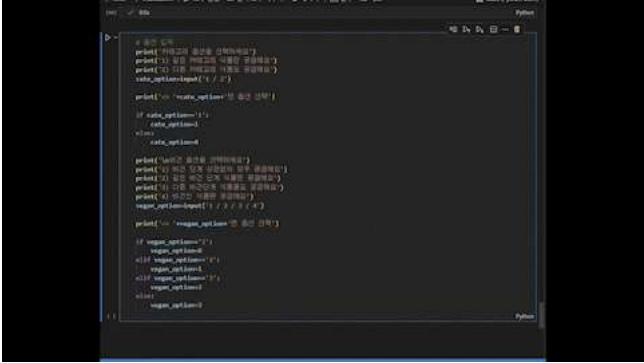
5. 추천시스템

입력받은 성분들과 비슷한 성분을 가진 제품을 추천하기에 앞서 사용자가 원하는 옵션을 먼저 입력받는다.

카테고리 옵션을 선택하세요
1) 같은 카테고리 식품만 공급해요
2) 다른 카테고리 식품도 공급해요
=> 1번 옵션 선택

비건 옵션을 선택하세요
1) 비건 단계 상관없이 모두 공급해요
2) 같은 비건 단계 식품만 공급해요
3) 다른 비건단계 식품들도 공급해요
4) 비건인 식품만 공급해요
=> 3번 옵션 선택

마켓컬리 사용예시 추천시스템 옵션 입력



마켓컬리 사용예시 추천시스템 옵션 입력 gif

사용자의 선택 옵션에 맞는 추천 상품들을 출력해 준다.

brand	item	ingredient	category	classify	allergy	gluten
1875	오리온 메감 치즈 그라탕 204g	롤란드산 산도조절제 식물성유지	간식/과자/떡	lacto	밀,우유,대두,쇠고기,오징어,새우	발견사항 없음
1570	칠갑농산 우리쌀 먹국떡 800g	우유 대두 계란	간식/과자/떡	lacto_ovo	난류(계란),우유,대두,밀	발견사항 없음
1593	오리온 고래밥 20g x 10입	백설탕 치즈분말 산도조절제 유화제 전분 갈색설탕 아스파탐 분말	간식/과자/떡	lacto	밀,달걀,우유,대두,토마토,돼지고기,쇠고기,오징어,조개류(굴)	발견사항 없음
2901	Gnaw 드라큐브 100g	전분	간식/과자/떡	vegan	우유,대두,밀,땅콩,호박가능성있음	발견사항 없음
2915	푸드드스드 페스츄리 한입약국 160g	대두	간식/과자/떡	vegan	밀,대두	발견사항 없음
933	서울마님 찰떡용심아	우유 대두 아황산류 새우 메밀엿갈	간식/과자/떡	pesco	알류,우유,메밀,땅콩,대두,밀,고등어,게,새우,돼지고기,복숭아,토마토,아황산류,호두...	발견사항 없음
1034	농심 자갈치 300g (치미떡)	우유 산도조절제 전분 과자 대두 명종 토마토 새우 미식고기 소맥분 당류 증합포화	간식/과자/떡	pesco	새우,대두,우유,밀,쇠고기,게,토마토,돼지고기,오징어,닭고기,계란,땅콩,맛조개류(L)	발견사항 없음
1677	해마루 파주 참단골 쌀소떡 (5개입)	설탕 밀가루 산도조절제 산도조절제전분 변성전분 맛술 베이킹파우더 대두 말조간장 땅콩...	간식/과자/떡	ovo	계란,밀,대두	밀가루
1730	대요리 로타 마리에 골드 초코 비스킷 2개입 빈들	밀가루 설탕 황색소금 시럽 베타 감고에 식물성유지 탄산수소암모늄 황색계면활성제	간식/과자/떡	lacto_ovo	밀,우유,계란,대두,견과류	밀가루
1523	오리온 초코파이 배나나맛 12입	변성전분 메치 천연색 식물성크립 전분 혼합제 향성향료 당류	간식/과자/떡	ovo	밀,달걀,우유,대두,돼지고기,쇠고기,땅콩,복숭아,게,조개류	발견사항 없음

마켓컬리 사용예시 추천시스템 결과 데이터프레임

b. 동식물성 화장품 분류 및 추천 시스템 구현

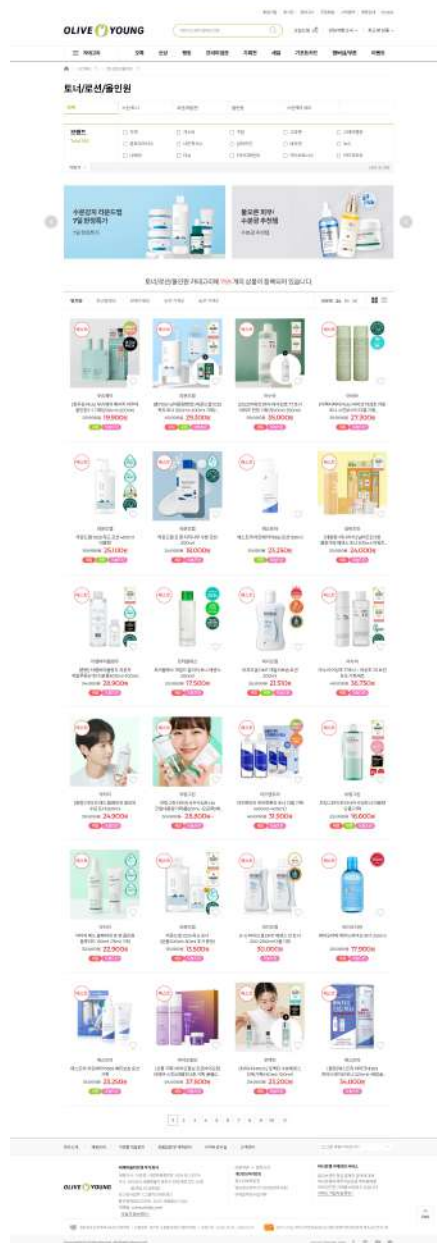
i. 데이터수집

올리브영 사이트에서 자체적으로 분류한 카테고리들 중 뷰티 탭에서 스킨케어, 마스크팩, 클렌징, 선크어, 메이크업, 네일, 바디케어, 헤어케어, 향수/디퓨저, 남성 카테고리를 선택하여 크롤링을 진행하기로 하였다.

뷰티		
스킨케어 토너/로션/올인원 에센스/크림 미스트/오일 마스크팩 시트팩/패드 페이셜팩 코팩/패치 클렌징 클렌징폼/젤 오일/워터/리무버 필링/패드 선크어 선블록 태닝/에프터션	더모 코스메틱 스킨케어 클렌징 선크어 마스크팩 바디케어 메이크업 립메이크업 베이스메이크업 아이메이크업 네일 폴리쉬 팁/스티커 반경화 케어 바디케어 워시/스크럽 로션/오일 핸드케어 립케어 제모용품 바디미스트 데오도란트 풋케어 선물세트	헤어케어 샴푸/린스 트리트먼트/팩 염색약/펌 헤어기기 스타일링 에센스 헤어브러쉬 탈모케어 향수/디퓨저 여성향수 남성향수 홈 프래그런스 선물세트 미용소품 페이스 아이 헤어/바디 코튼 디바이스 네일 기타소품 남성 스킨케어 헤어케어 쉐이빙 향수/메너용품 메이크업 바디케어

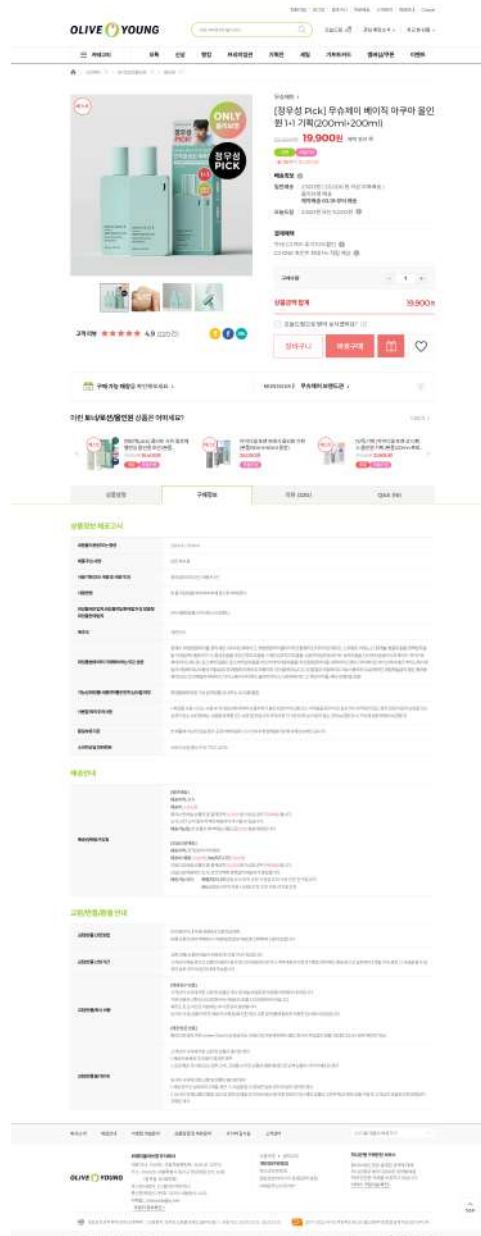
올리브영 카테고리 종류

소카테고리로 들어가면 한 페이지당 24개의 상품을 보여주고 있고, 다음 페이지로 넘어가는 숫자 버튼이 존재했다.



올리브영 소카테고리 세부페이지

상품을 클릭하여 상품페이지로 들어가면 브랜드, 상품명, 가격 등의 정보가 존재했다. 우리가 필요로 하는 성분 정보는 구매정보 탭을 클릭하면 확인할 수 있었다.



올리브영 상품 상세페이지

원하는 정보들의 위치를 파악해 봤을 때, 소카테고리 상세페이지에서 한 상품씩 클릭하여 상품페이지로 이동 후, 브랜드 이름과 상품 이름을 크롤링 해오고, 구매정보 탭을 클릭하여 화장품 성분을 크롤링 한 후 이전 페이지로 돌아가, 다음 상품을 클릭하는 과정을 반복하는 코드를 작성해 주었다. 한 페이지에 있는 24개의 상품을 모두 크롤링을 했다면, 다음 페이지로 넘어가는 숫자 버튼을 찾아 클릭한 후 다음 페이지에서도 전과 똑같은 작업을 반복해 주었다.

정적 크롤링과 동적 크롤링을 반복해서 사용해 주기 때문에 BeautifulSoup과 Selenium을 같이 사용해 주며 코드를 작성하였다.

```
# 예시 - 여성향수
url='https://www.oliveyoung.co.kr/store/display/getMCategoryList.do?dispCatNo=100000100050003&isLoginCnt=3&aShowCnt=0&bShowCnt=0&cShowCnt=0'
driver=webdriver.Chrome()
driver.get(url)
act=ActionChains(driver)

text_lst=[]
brand_name=[]
product_name=[]
page=0
for page in range(0,9): # 페이지 for문
    for line in range(8,14): # 행 for문
        for i in range(4):
            # 상품 클릭
            driver.find_elements(By.CSS_SELECTOR, '#Contents > ul:nth-child('+str(line)+') > li')[i].click()
```

```

time.sleep(0.5)

# 브랜드 이름, 상품 이름 크롤링
html=driver.page_source
soup=BeautifulSoup(html, 'html.parser')
brand_name.append(soup.select('div.prn_info > p')[0].text)
product_name.append(soup.select('div.prn_info > p')[1].text)

# 구매정보 클릭 및 성분 크롤링
driver.find_elements(By.CSS_SELECTOR, '#buyInfo')[0].click()
time.sleep(0.5)
try:
    html=driver.page_source
    soup=BeautifulSoup(html, 'html.parser')
    text_lst.append(soup.select('div > dl > dd')[8])
except:
    text_lst.append('')

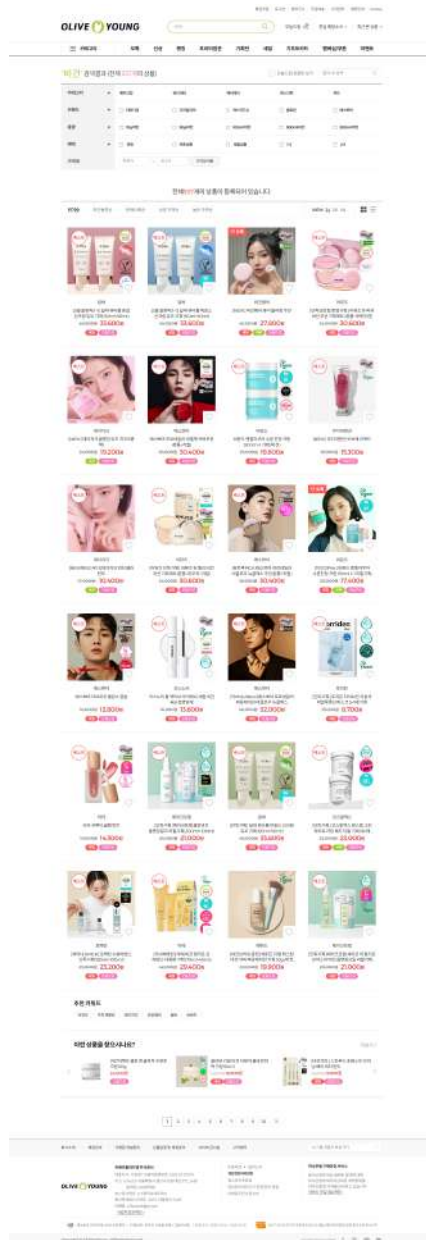
# 이전 화면으로 이동
driver.back()

# 다음 페이지로 이동
time.sleep(0.5)
driver.find_elements(By.CSS_SELECTOR, '#Container > div.paging > a')[page].click()
print('page', page)

```

이와 같은 작업을 소카테고리마다 url를 새로 지정해주며 위 작업을 반복하였다.

올리브영 검색창에 '비건'이라고 검색하면 비건인 상품들을 보여준다. 소카테고리와 마찬가지로 기본으로 24개의 상품씩 보여준다.



올리브영 비건 검색 후 페이지

우리가 이전에 크롤링 한 상품들 중에서 비건인 상품은 무엇이 있는지 확인하기 위해 브랜드 이름과 상품 이름이 필요했다. 따라서 상품 상세페이지로는 들어가지 않고 소카테고리 페이지에서 브랜드 이름과 상품 이름을 크롤링 해오는 작업을 진행하였다.

```
brand_lst=[]
product_lst=[]

for url in url_lst:
    driver=webdriver.Chrome()
    driver.get(url)
    act=ActionChains(driver)

    html=driver.page_source
    soup=BeautifulSoup(html, 'html.parser')

    for i in range(24):
        brand_lst.append(soup.select('div.prd_info > div.prd_name > a > span')[i].text)
        product_lst.append(soup.select('div.prd_info > div.prd_name > a > p')[i].text)
```

수집한 비건 데이터 중에 이전에 수집한 올리브영 데이터에 존재하는 상품들에 한해서 합하여 최종 데이터 셋을 완성해 csv파일로 저장하였다.

	brand	product	text	cate	vegan
0	\n포맨트\n	[2월 올영픽] 포맨트 시그니처 퍼퓸 코돈허그 50ml 기획(+핸드크림 30ml 증정)	[에탄올, 향료, 정제수, 피이지-60하이드로제네이티드캐스터오일, 부틸렌글라이콜, ...	남성	0
1	\n랑방\n	[2월 올영픽] 랑방 에끌라 드 아르페쥬 EDP 30ml(+2ml 샘플*2ea)	[변성알코올39-C, 정제수,향료, 부틸페닐메틸프로피오날, 리모넨, 시트랄, 리날롤...	남성	0
2	\n포맨트\n	[2월 올영픽] 포맨트 코튼메모리 퍼퓸 50ml 기획(+핸드크림 30ml 증정)	[에탄올, 향료, 정제수, 피이지-60하이드로제네이티드캐스터오일, 부틸렌글라이콜, ...	남성	0
3	\n포맨트\n	포맨트 시그니처 퍼퓸 50ml (코튼디어나잇)	[에탄올, 향료, 정제수, 피이지-60하이드로제네이티드캐스터오일, 부틸렌글라이콜, ...	남성	0
4	\n클린\n	[2월 올영픽] 클린 워크톤 30ml 기획(클러볼 10ml 증정)	[* 클린 워크톤 - 변성알코올,향료,정제수,글리세린,알로에베라잎즙,페녹시아에탄올,벤...	남성	0

올리브영 데이터 크롤링 후 데이터 프레임

ii. 데이터 전처리

전처리가 필요한 부분들은 살펴보자.

brand	product	text
\n포맨트\n	[2월 올영픽] 포맨트 시그니처 퍼퓸 코돈허그 50ml 기획(+핸드크림 30ml 증정)	[에탄올, 향료, 정제수, 피이지-60하이드로제네이티드캐스터오일, 부틸렌글라이콜, ...
\n포맨트\n	[2월 올영픽] 포맨트 시그니처 퍼퓸(벨벳허그) 50ml+전용 파우치 증정	[변성알코올, 향료, 정제수, 피이지-40하이드로제네이티드캐스터오일, 알파-아이소메...
\n루아페\n	[2/14 하루특가] 루아페 디어랑스 손리드 퍼퓸 2종 택 1	[513] 사이클로헥산티올세렌, 사이클로헥사실세렌,다이메틸콘크로스폴리메, 글리세...
\n랑방\n	[2월 올영픽] 랑방 에끌라 드 아르페쥬 EDP 30ml(+2ml 샘플*2ea)	[변성알코올39-C, 정제수,향료, 부틸페닐메틸프로피오날, 리모넨, 시트랄, 리날롤...
\n클린\n	[2월 올영픽] 클린 워크톤 30ml 기획(클러볼 10ml 증정)	[* 클린 워크톤 - 변성알코올,향료,정제수,글리세린,알로에베라잎즙,페녹시아에탄올,벤...
...
\n돈28\n	돈28 고체향수 2종 택1 (들은퍼퓸 9ml+파우치 증정)	[1. 화장품 사용 시 또는 사용 후 적자 광선에 의하여 사용부위가 붉은 반점, ...
\n에스\n	[NEW] 에스 비터 스윗 EDP 60ml	[에탄올,향료,글리세린,아들라시시다목부오일,에칠헥실메톡시신나메이트,메틸프로판다이올...
\n겐조\n	[NEW] 라 키텍션 겐조 메모리 레브 모두스 EDP 75ml	[1. 화장품 사용 시 또는 사용 후 적자광선에 의하여 사용부위가 붉은 반점, 부어...
\n에스\n	[NEW] 에스 워비 포레스트 EDP 60ml	[에탄올, 글리세린, 향료,자몽껍질오일,에칠헥실메톡시신나메이트,메틸프로판다이올,커먼...
\n페라가모\n	페라가모 세노리니 미스테리오시 EDP 50ml	[변성알코올39-C,향료,정제수,리모넨,벤질살리실레이트,리날롤,에칠헥실메톡시신나메이...

올리브영 데이터 전처리 전 데이터프레임

[brand, product 전처리]

먼저 brand에서는 브랜드 이름과 함께 '\n'이 등장하여 브랜드 이름만 살리고 나머지는 제거해 주는 작업이 필요했다. 또한 product에서는 [] 안에 '2월 올영픽', 'NEW', '하루 특가' 와 같은 상품 이름과 관련 없는 정보들을 제거해 줘야 했다. 이는 정규 표현식을 사용하여 필요한 부분만 살리고 제거해 주었다.

```
brand=[]
product=[]

for b in brands:
    brand.append(re.sub('[\n]', '', b))

for p in products:
    try:
        while(1):
            idx=re.search('[\n]',p).end()
            p=p[idx:]
    except:
        pass
    product.append(p.strip())
```

[성분 전처리]

메이크업 도구이거나, 보통의 다른 상품들과 다른 곳에 성분들이 적혀있는 상품들이 몇 개 존재하였다. 이와 같은 상품들은 성분 대신 화장품 사용 시 주의사항, 교환 관련 안내문 등이 크롤링 된 것을 확인할 수 있었다. 따라서 이와 데이터는 제거해 주기로 결정하였다.

```
data=data[~data['text'].str.contains('1\')]
data=data[~data['text'].str.contains('화장품')]
data=data[~data['text'].str.contains('교환')]
data=data[~data['text'].str.contains('사용')]
```

이후 성분에 대한 정보만 남기기 위해 데이터를 살펴본 결과 다음과 같은 특징을 찾을 수 있었다.

1. 크롤링을 한 데이터이다 보니 <dd>,
과 같은 HTML 태그들이 포함되어 있었다.

2. 여러 상품이 세트인 상품들은 성분 데이터 안에 상품 이름이 함께 적혀있는 것이 있었다.
3. 다양한 특수기호가 포함되어 있었다.
4. 상품 이름이 [] 안에 존재하기도 했고, ;, - 와 같은 기호들로 상품 이름과 성분들을 나누는 경우도 있었다.
5. 성분을 나열하는 방식에는 ,로 나열하는 방법과 띄어쓰기로 나열하는 방법이 존재했다.

우선 성분들을 나열하는 방식을 콤마(,)로 통일시켜 주었다.

이후 [] 안에 존재하는 내용을 []와 함께 삭제해 주었고, 다양한 특수기호와 HTML 태그로 성분들을 구분하는 방식에서 콤마(,)로 구분하는 방법으로 통일시켜 주었다.

```
lst=[]
for text in text_lst:
    # 성분 하나씩 확인하기
    if '@' in text:
        text=text.split('@')
    elif ',' not in text:
        text=text.split(' ')
    else:
        text=text.split(',')

    result=''
    for word in text:
        # [] 내용 제거
        cnt=0
        try:
            while(cnt<3):
                if cnt>3:
                    break
                start_idx=re.search('\[',word).start()
                end_idx=re.search('\]',word).end()
                word=word[start_idx:word[end_idx:]]
                cnt+=1
        except:
            pass

        # 특수기호들 ,로 통일하기
        word=word.strip()
        word=word.replace(' ','')
        word=word.replace('-',',')
        word=word.replace(':',',')
        word=word.replace('<dd>','')
        word=word.replace('</dd>','')
        word=word.replace('■','')
        word=word.replace('<br/><br/>','')
        word=word.replace('<br/>','')
        word=word.replace('/',',')
        word=re.sub('[\t]',' ',word)
        result=result+word+' '
    lst.append(result)
```

이후 다시 성분 데이터를 보며 해결되지 못한 부분들을 찾아보았다. 주로 여러 상품을 한 패키지로 묶어서 판매하는 상품들에서 찾을 수 있는 부분들이었다.

1. 성분 중 정제수는 거의 첫 번째로 등장하는 성분이다. 상품명과 정제수가 붙어있는 경우 위의 전처리 과정에서 처리되지 않았기 때문에 따로 처리해 주는 작업이 필요했다.
2. 본품 구성 00로션, 00토너 00ml 와 같이 상품의 이름이 위의 전처리 과정에서 제거되지 않은 경우가 발견되었다.

1번의 경우에는 정제수라는 단어가 포함되어 있는 경우 정제수만 살리고 함께 있는 단어는 삭제해 주는 작업을, 2번의 경우 본품, 로션, 토너, ml과 같은 단어가 포함되어 있는 경우 그 단어를 아예 삭제해 주는 작업을 진행해 주기로 하였다.

```
result=[]
for text in lst:
    text_lst=[]
    for word in text.split(','):
        if word=='':
            pass

        if '정제수' in word and len(word)>4:
            word='정제수'

        if '올인원' in word and len(word)>4:
            word=' '
```



```

if '토너' in word and len(word)>3:
    word=''

if '포마드' in word and len(word)>4:
    word=''

if '로션' in word and len(word)>3:
    word=''

if 'мл' in word:
    word=''

if '본품' in word:
    word=''

if len(word)>0:
    text_lst.append(word)
result.append(', '.join(text_lst))

```

여러 상품이 담긴 패키지 상품일 경우 같은 성분이 중복해서 등장할 수 있기 때문에 성분들이 중복 없이 하나씩 존재할 수 있도록 처리해 주었다.

```

# 중복 성분 제거
result=[]
for lst in data['text']:
    text=lst.split(',')
    text=list(set(text))
    result.append(', '.join(text))

```

올리브영의 홈페이지의 경우 시간이 흐름에 따라 화면을 구성하고 있는 상품들의 종류가 변하게 된다. 크롤링이 여러 시간에 걸쳐 진행되었기 때문에 크롤링 한 데이터 중에서 중복되는 항목이 있을 것이라고 판단하여 브랜드와 상품명이 중복되는 항목들은 삭제해 주었다.

```

# 중복행 제거
data=data.drop_duplicates(['brand', 'product'], keep='first')
data

```

	brand	product	text	cate	vegan
0	아이디얼 포맨	아이디얼 포맨 프레시 올원원 기왈 (분홍150ml+50ml 중량)	베타글루칸, 소듐하이드로록사이드, 다이옥살리판올하이드로록사이드, 녹차수, 아티초크잎추출물, ...	스킨케어	0
1	아누아	아누아 아성초 77 토너 에워즈 한정 기획 (500ml+250ml)	마트리카리아꽃추출물, 시달수수추출물, 아이소벤일다이올, 체이스트리추출물, 글리세린, 다이소...	스킨케어	0
2	넵버즈린 3번	넵버즈린 3번 결광가득 에센스 토너 300ml 에워즈 한정기획 (+3번 결광케어 케...	바실러스, 보리씨발효여과물, 조류추출발효물, 홍삼발효여과물, 쌀겨발효여과추출물, 나이아신아마...	스킨케어	0
3	에스트라	에스트라 아도베리어365 에워즈 보습 로션 기획	비타민이소스테라이드, 글라스테롤, 팔메이트, 부틸렌글라이콜, 디카프알레이트, 글루코오스디스...	스킨케어	0
4	더럽버어분광두	더럽버어분광두 올리고 리얼루문산 5000토너 200ml+200ml 리필기획	다이옥살리판올하이드로록사이드, 베타글루칸, 동백나무잎추출물, 에틸헥실글리세린, 500ppm, 알란토인...	스킨케어	0
...
3748	옛지유	옛지유 섬케어 담쟁 글로시&메트	유칼립투스잎추출물, 아이드록사이드, 글라스테롤, 셀룰로오스아세테이트, 부티레이트, 헤마, 메...	네일	0
3749	대심디바	대심디바 코어 세럼	박아오린, 변성알코올, 프로판올, 글라이콜, 정제수, 디메틸실리콘, 키토산	네일	0
3750	대심디바	대심디바 리치 세럼	벤질살리실레이트, 토드씨오일, 미네랄오일, 스타이렌코폴리머, 부틸렌, 에틸렌, 부틸렌디메틸프로...	네일	0
3751	위드산	위드산 에코 네일 리무버 200ml	제라나올, 리날올, 청색1호(CI42090), 프로판올, 글라이콜, 정제수, 라벤더오일, 적색22...	네일	0
3752	위드산	위드산 메가사안단코트 15ml	아디픽에씨드, 에틸아세테이트, 자색201호, 트리멜리틱안하이드라이드코폴리머, 에탄올, 나이트...	네일	0

올리브영 데이터 일부 전처리 후 데이터프레임

이후 데이터를 살펴봤을 때 괄호가 포함되어 있는 것을 확인할 수 있었다. 이후 작업을 생각해 봤을 때 괄호 없이 성분들로만 나열되어 있는 것이 옳다고 판단되어 괄호는 제거해 주고 안에 괄호 안의 내용을 살려두는 처리를 해주었다.

```

# 괄호제거
def hasNumber(stringVal):
    return any(elem.isdigit() for elem in stringVal)

text_lst=[]
for text in df['text'].tolist():
    sen=''
    for word in text.split(','):
        if '(' in word:
            word=word.split('(')[1]
        if ')' in word:
            word=word.split(')')[0]
        if len(word)>0 and hasNumber(word)==0:

```

```

sen=sen+', '+word
sen=sen[1:]
text_lst.append(sen)

df['text']=text_lst

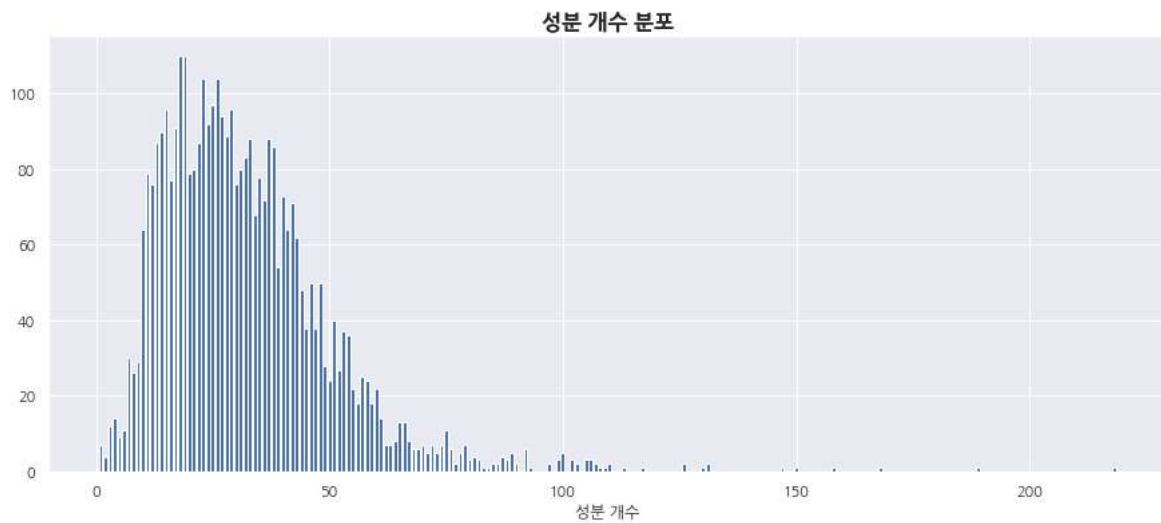
```

이렇게 전처리를 마친 총 3753개의 올리브영 데이터를 csv파일로 저장하였다.

	brand	product	text	cate	vegan
0	아이디얼포맨	아이디얼 포맨 프레스시 올인원 기믹 (본품150ml+50ml 증정)	베타글루칸,소듐하이알루로네이트,다이메틸실란올하이알루로네이트,녹차수,아티초크잎추출물...	스킨케어	0.0
1	야누아	야누아 여성초 77 토너 여워즈 한정 기믹 (500ml+250ml)	미트리카리아꽃추출물,사탕수수추출물,아이소벤질디아올,체이스트트리추출물,글리세린,다이소...	스킨케어	0.0
2	넵바즈민	넵바즈민 3인1용 걸광가득 에센스 토너 300ml 여워즈 한정 기믹 (+3인1용 걸광케어...	바실러스,보리씨발효여과물,조류수출발효물,홍삼발효여과물,참기발효여과수출물,나이아신아...	스킨케어	0.0
3	에스트라	에스트라 아토페라어365 예민보습 로션 기믹	베한이소스테아레이트,글래스탈론,알메락메씨드,부딕헨글라이콜디카프릴레이트,글루코오스...	스킨케어	0.0
4	더럼바이블링두	더럼바이블링두 올리고 히알루론산 5000도너 200ml+200ml 리필기믹	다이소듐이디티에이,베타글루칸,중쇄나우알수출물,에틸헥실글리세린,알란토인,정제수,판테...	스킨케어	0.0
...
3748	옛지유	옛지유 샴페어 답쟁 글로시&메트	유칼립투스잎추출물,하이드록시사이클로헥실페닐케톤,셀룰로오스아세테이트부타디엔,헤...	네일	0.0
3749	데성디바	데성디바 코어 세럼	바이오틴,변성알코올,프로판렌글라이콜,정제수,디메틸설펜,키토산	네일	0.0
3750	데성디바	데성디바 러치 세럼	벤질살리실레이트,포도씨오일,미네랄오일,스타이렌코폴리머,부딕헨,에틸렌,부틸페닐메틸프...	네일	0.0
3751	위드산	위드산 에코 네일 리무버 200ml	제라니올,리날룰,프로필렌글라이콜,정제수,라벤더오일,토코페릴아세테이트,오레가노오일,코...	네일	0.0
3752	위드산	위드산 메가사인암코트 15ml	아디픽메씨드,에틸아세테이트,트리ethyl안하이드라이드코폴리머,에탄올,나이트로셀룰로오...	네일	0.0

올리브영 데이터 전처리 완료 데이터 프레임

한 상품당 얼마나 많은 성분들을 포함하고 있는지 살펴보기 위해 상품당 포함 성분 개수 분포를 그려보았다. 대부분 한 상품당 50개 미만의 성분들을 포함하고 있는 것을 확인할 수 있다.



올리브영 상품당 포함 성분 개수 분포

iii. 분류 시스템 구현

올리브영 데이터를 2가지 종류로 분류하는 시스템을 구현하고자 한다. 성분 데이터를 통해 동물성 성분이 하나라도 포함되어 있다면 동물성 화장품으로, 동물성 성분이 하나도 포함되지 않다면 식물성 화장품으로 분류할 것이다.

이 작업을 하기 위해서는 우선 동물성 성분 리스트가 필요하다. 이전 작업을 통해 쌓아둔 동물성 성분 리스트(animal.txt)를 불러와 animal_lst에 저장하였다.

```

# 동물카테고리 불러오기
with open("./list/animal.txt", "r") as file:
    lst = file.readlines()

```

```

animal_lst=[]
for word in lst:
    word=word.replace('\n','')
    word=word.replace(' ','')
    if len(word)>0:
        animal_lst.append(word)

```

이후 크롤링을 통해 구한 비건 화장품들의 성분들을 추출하여 vegan_lst를 만들었다.

```

# 비건카테고리 성분리스트
vegan_lst=[]
for sen in df.loc[df['vegan']==1.0,'text'].tolist():
    for word in sen.split(','):
        vegan_lst.append(word)
vegan_lst=list(set(vegan_lst))

```

비건 카테고리에서 뽑아낸 성분 리스트에는 동물성 성분이 포함되지 않았기 때문에 혹시 animal_lst에서 vegan_lst에 해당하는 성분이 있다면 제거해주어 올리브영 화장품을 동물성 화장품과 식물성 화장품으로 분류할 때 사용할 최종 체크리스트(check_lst)를 완성시켰다.

```

# 분류 체크리스트
check_lst=list(set(animal_lst)-set(vegan_lst))
check_lst

```

coma(,)를 통해 나열되어 있는 성분들을 하나씩 확인해가며 만약 한 성분이라도 체크리스트에 포함되어 있다면 동물성 성분이 들어간 것으로 판단하여 동물성 화장품으로, 체크리스트를 통과하며 하나라도 걸리지 않았다면 식물성 화장품으로 분류하는 함수를 만들어 분류를 진행하였다.

```

# 동/식물 분류
def classify(lst):
    for word in lst.split(','):
        word=word.replace(' ','')
        if word in check_lst:
            ingredient_lst.append(word)
            return 'animal'

ingredient_lst=[]
df['class']=df['text'].apply(classify)
df['class'].fillna('not animal',inplace=True)

```

	brand	product	text	cate	animal
0	렙시리즈	렙시리즈 울인원 트리트먼트 50ml 1+1 기희세트	사이클로메트릭스,포타슘스테아레이트,포도씨추출물,스쿠알란,옥틸도데실테오코넨타노에이트,메...	스킨케어	1
1	유세인	유세인 하이알루론 아이크림x나이드크림 기희(아이크림 15ml+ 나이드크림 50ml+ 컨...	에칠헤실리세인,부틸렌글리콜,디카프릴레이트,아소프로필팔미테이트,하이드로제네이티드코...	스킨케어	1
2	아이디얼포맨	아이디얼 포맨 퍼펙트스킨케어 2중세트(미니여처 3중 중점)	페닐알라닌,무화과추출물,폴리비질릴추출물,나이에신아마이드,가프릴락,다이스도이디티에이...	스킨케어	1
3	다슈	다슈 맨즈 야쿠아 토너/로션 153ml 2중 세트(+ 토너&로션&글관정 30ml중점)	개황각추출물,연꽃씨추출물,바실러스,마드라카리아꽃추출물,피루리발효추출물,페닐알라닌...	스킨케어	1
4	달바	달바 화이트 드러플 퍼스트 아로마틱 토너 155ml	테오펜알글라이콜다이하텐타노에이트,하이드록시에틸우레아,등공오일,브로멜라인,소듐글루...	스킨케어	1
...
3748	대상디바	대상디바 메직프레스 베이스 쉼드	토코페롤,스쿼트아몬드오일,아이스프로필알코올,해바라기씨오일,포스포릭에씨드,실리카,트...	네일	0
3749	엣지유	엣지유 삼계어 담젤 글로시&매트	유칼립투스잎수출물,히이드록시사이클로헥실페닐케톤,셀룰로오스아세테이트부티레이트,해...	네일	0
3750	대상디바	대상디바 리치 세럼	벤질살리실레이트,포도씨오일,미네랄오일,스타이렌코폴리머,부틸렌,에틸렌,부틸페닐메틸프...	네일	0
3751	위드산	위드산 예코 네일 리무버 200ml	제라나올,리날올,프로필렌글라이콜,정제수,라벤더오일,토코페릴아세테이트,오레가노오일...	네일	0
3752	위드산	위드산 메카샤먼탑코트 15ml	아디픽에서드,에틸아세테이트,트리멜리틱안하이드라이드코폴리머,에탄올,나이트로셀룰로...	네일	0

3753 rows × 6 columns

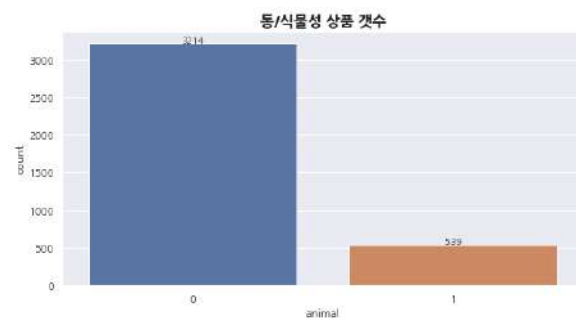
분류가 완료된 올리브영 데이터 프레임

	brand	product	text	cate	animal
9	템시리즈	템시리즈 울연원 트리트먼트 50ml 1+1 기증세트	사이클로덱스트린,포타슘스테아레이트,포도씨추출물,스쿠알란,옥틸도데실테오코넬타노에이트,메...	스킨케어	1
15	유세권	유세권 라이알루론 아이크림X나이트크림 기증(아이크림 15ml+나이트크림 50ml+ 린...	에틸헥실리세린,부틸렌글리콜,디카프릴레이트,이소프로필알마테이트,라이도르제나이트,디코코리세...	스킨케어	1
29	아이디얼포맨	아이디얼 포맨 퍼펙트스킨케어 2중세트(마니여저 3중 증정)	페닐알라닌,무화과추출물,올리브잎추출물,나이아신아마이드,카프릴릭,다이소스테로이드,...	스킨케어	1
34	다슈	다슈 맨즈 아쿠아 토너/로션 153ml 2중 세트(+ 토너&로션&클렌징 30ml증정)	개장각추출물,연꽃씨추출물,비살라스,마트리카리아꽃추출물,피부리알로추출물,페닐알라닌,민...	스킨케어	1
38	달바	달바 화이트 트러를 파스트 아로마틱 토너 155ml	네오펜틸글리콜,다이하이드라노에이트,화이트독시메틸우레아,들금오잎,프로필라민,소듐글루타이드...	스킨케어	1
...
3597	줄꽃	줄꽃 쏫 리퍼어 크림 120ml	타일알주출물,세테아일알코올,소듐파타레이트,하이드록시아세토페논,식물성스쿠알란,라벤더꽃...	바디	1
3604	티타니아	티타니아 크림드림밤	소듐벤조에이트,토코페롤,소듐스테아로일라일레이트,세테아일알코올,라릭에릭트,페녹시아판올...	바디	1
3606	소프리스	소프리스 쏫 더블에센스 마스크	스테아릭에릭트,폴리글루타릭에씨드,하이드록시시트로넬알,부틸벤질메틸프로피오날,빈도올리고...	바디	1
3729	코스노리	코스노리 실크리퍼어 네일 크림 손톱영양제	브이피코폴리머,미리스틸알코올,에소스테아릴,발거오일,세틸알코올,세테아일알코올,하이드록...	네일	1
3749	대성디바	대성디바 코어 세럼	네오오린,벤성알코올,프로필렌글리콜,정제수,디메틸실리콘,기토산	네일	1

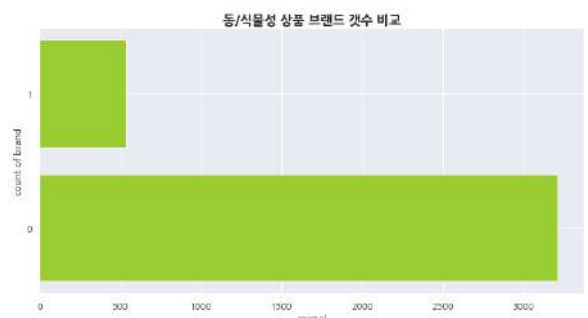
동물성 화장품 데이터프레임

	brand	product	text	cate	animal
0	아이디얼포맨	아이디얼 포맨 프레시 울연원 기증 (본품150ml+50ml 증정)	베타글루칸,소듐하이알루로네이트,다이하이드라노에이트,나이아신아마이드,녹차수,아티초크잎추출물...	스킨케어	0
1	여누마	여누마 여성조 77 토너 어워즈 한정 기증 (500ml+250ml)	마트리카리아꽃추출물,사탕수수추출물,아이소헥실다이올,제아스트리크추출물,글리세린,다이소...	스킨케어	0
2	넵비즈인	넵비즈인 3번 걸광기증 에센스 토너 300ml 어워즈 한정 기증 (+3번 걸광케어 기...	비살라스,보리씨발효여과물,조류추출발효물,홍삼발효여과물,발거발효여과추출물,나이아신아마...	스킨케어	0
3	엑스트라	엑스트라 아도베리아365 메이크업 로션 기증	비탄아소스테아레이트,글라세틴,발매릭에릭트,부틸렌글리콜,디카프릴레이트,글루코코사이드...	스킨케어	0
4	더립베이블링두	더립베이블링두 울리고 히알루론산 5000토너 200ml+200ml 리얼기증	다이소스테로이드,네타글루칸,동백나무잎추출물,에틸헥실글리세린,알란토인,정제수,관대...	스킨케어	0
...
3747	대성디바	대성디바 메이크프레스 베이스 실크	토코페롤,스위트아몬드오일,아이소프로필알코올,헥바리기씨오일,포스포리페이드,실리카,드라...	네일	0
3748	엠티유	엠티유 상케어 입점 글로시&메드	유알렌트스알주출물,하이드록시사이클로헥실페닐케톤,폴로모오스테아레이트부티레이트,헤마...	네일	0
3750	대성디바	대성디바 리치 세럼	벤질살리실레이트,포도씨오일,미네랄오일,스타이렌코폴리머,부틸렌 아릴렌,부틸페닐메틸프로...	네일	0
3751	위드산	위드산 에코 네일 리무버 200ml	저라니올,리날롤,프로필렌글리콜,정제수,라벤더오일,토코페릴아세테이트,오레가노오일,로...	네일	0
3752	위드산	위드산 메가사인팅코트 15ml	아디픽에릭트,메틸아세테이트,트리메틸아민하이드라이드코폴리머,에탄올,나이트로셀룰로오스...	네일	0

식물성 화장품 데이터프레임



올리브영 분류 데이터 결과 시각화 - 동/식물성 상품 갯수 비교



올리브영 분류 데이터 결과 시각화 - 동/식물성 상품 브랜드 갯수 비교

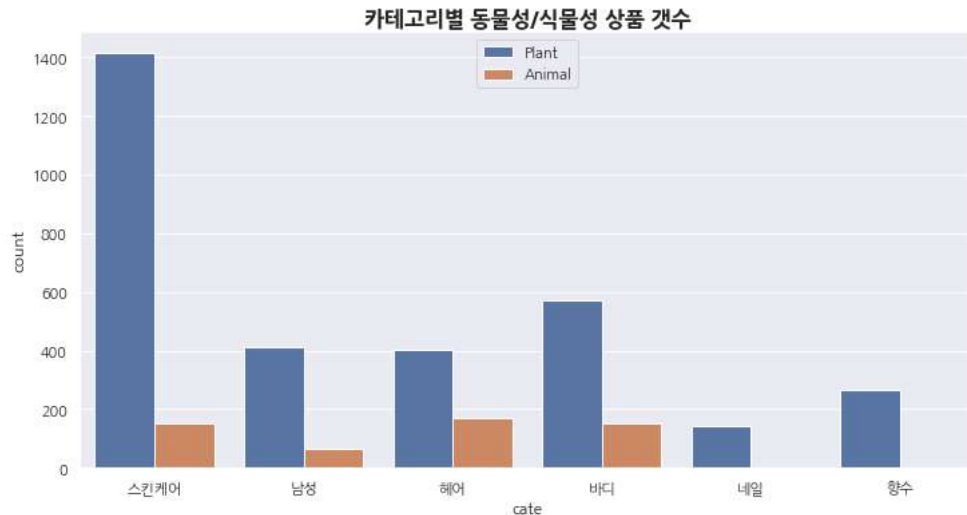
그 결과 올리브영 데이터를 539개의 동물성 화장품과 3214개의 식물성 화장품으로 분류된 것을 확인할 수 있다.

동물성 화장품을 판매하고 있는 브랜드의 수도 식물성 화장품을 판매하고 있는 브랜드의 약 6배에 해당하는 것도 확인할 수 있다.

대부분의 화장품이 동물성 화장품으로 분류될 것이라고 예상했던 것과 달리 대부분의 화장품이 식물성 화장품으로 분류된 것을 확인할 수 있다.

보통 비건 화장품이란 화장품 제조 가공 단계에서 동물성 원료를 일절 사용하지 않고 동물실험을 하지 않는 화장품을 말한다. 우리가 분류를 할 때 사용한 기준이 되는 성분만으로는 동물실험 여부는 확인할 수 없었고, 동물 실험의 여부는 규정상 따로 언급을 하지 않는 이상 회사에 직접 문의를 통해 알아내는 방법뿐이라는 점에서 한계가 있었다. 또한 화장품은 대부분 화학성분들로 구성되어 있지만 우리는 동물성 성분의 포함여부만 파악했기 때문에 식물성 화장품이라고 분류됐지만 식물성 성분이 존재하지 않을 수도 있다. 하지만 성분만을 통해서 동물성 성분이 포함되었는지 포함되지 않았는지 판단하겠다는 우리의 주제를 생각해 봤을 때는 예상과는 달랐지만 원하는 결과물을 얻어낼 수 있었다고 판단하였다.

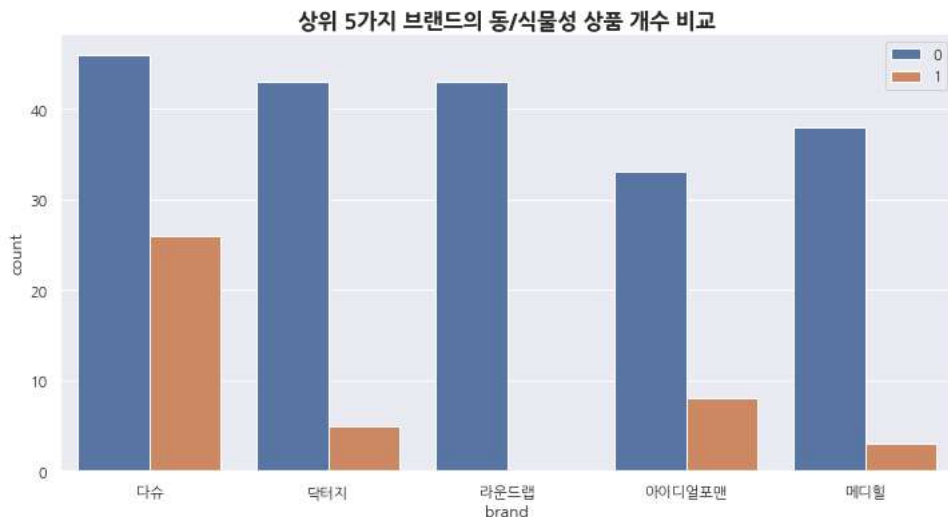
카테고리별로 동물성과 식물성 상품의 갯수를 비교해보았다.



올리브영 분류 데이터 결과 시각화 - 카테고리별 동식물성 상품 개수 비교

네일과 향수에서는 동물성 성분을 포함하고 있는 상품의 개수가 다른 카테고리에 비해 현저히 적은 것을 확인할 수 있었다. 이는 네일과 향수는 대부분 화학 성분들을 사용해서 만들기 때문에 나타난 결과라고 판단된다.

올리브영 데이터에서 가장 많이 등장한 5가지 브랜드를 뽑아 동물성 화장품과 식물성 화장품의 개수를 비교해보았다.



올리브영 분류 데이터 결과 시각화 - 상위5가지 브랜드 동식물성 상품 개수 비교

라운드랩에서는 동물성 성분이 포함된 화장품을 발견하지 못했다. 라운드랩이라는 브랜드는 자연 그대로의 깨끗한 원료를 사용한다는 브랜드 신념에 맞게 '자작나무 선크림', '소나무 진정 시카 로션' 등과 같은 화장품을 판매하고 있기 때문이라고 보인다.

iv. 추천 시스템 구현

전처리와 분류를 마친 올리브영 데이터를 불러와서 추천 시스템을 구현해 보았다.

	brand	product	text	cate	animal
0	럼시리즈	럼시리즈 플인원 트리트먼트 50ml 1+1 기워세트	사이클로헥스트린,포타슘스테아레이트,포도씨추출물,스쿠알란,옥틸도데실네오헥타노에이트,에...	스킨케어	1
1	유세린	유세린 허아일루온 아이크림x나이트크림 기워(아이크림 15ml+나이트크림 50ml+건...	에실렉실리세린,부틸렌글리콜,디카프릴레이트,이소프로필팔미테이트,하이드로제네이티드코코리세...	스킨케어	1
2	아이디얼포맨	아이디얼 포맨 피렉트스킨케어 2종세트(매니저저 3종 증정)	페닐알라닌,무화과추출물,홍리버칼릴추출물,나이아신아마이드,카프릴리,다이소듐이디티에이,...	스킨케어	1
3	다슈	다슈 멘즈 아쿠아 토너/로션 153ml 2종 세트(+ 토너&로션&클렌징 30ml증정)	계정각추출물,간꽃씨추출물,배실라스,매트리카리아올추출물,피루리알프추출물,페닐알라닌,민...	스킨케어	1
4	달바	달바 화이트 트러플 파스트 아로마틱 토너 155ml	네오칼릴글라이콜다이헥타노에이트,헥사드록시에틸우레아,글루코실,브로멜라인,소듐글루코라이드...	스킨케어	1
...
3748	대성디바	대성디바 맥직프레스 베이스 얼드	토코페롤,스위트아몬드오일,아이스프로필알코올,해바라기씨오일,포스포티에세드,실리카,트라...	네일	0
3749	옛지유	옛지유 삼케어 탭얼 글로시&메드	유칼립투스잎추출물,하이드록시사이클로헥실페닐케톤,셀룰로오스아세테이트부타레이트,해마,매...	네일	0
3750	대성디바	대성디바 리지 세럼	벤질살리실레이트,포도씨오일,미네랄오일,스타이렌코폴리머,무결렌,에틸렌,부틸페닐메틸프로...	네일	0
3751	위드산	위드산 예코 네일 러무버 200ml	제라니올,리날룰,프로필렌글라이콜,정제수,라벤다오일,토코페릴아세테이트,오레가노오일,로...	네일	0
3752	위드산	위드산 메가사인팅코트 15ml	아디픽에씨드,에틸아세테이트,트리멜리트산하이드라이드코폴리머,에탄올,나이트로셀룰로오스...	네일	0

올리브영 데이터 프레임

상품을 추천하는 기준은 화장품의 성분이다. 따라서 먼저 CountVectorizer를 사용하여 화장품의 성분들을 벡터화 시켜주었다.

```
from sklearn.feature_extraction.text import CountVectorizer

df['text']=df['text'].apply(lambda x: x.replace(',',' '))
count_vect=CountVectorizer(min_df=0,ngram_range=(1,2))
text_mat=count_vect.fit_transform(df['text'])
```

이후 코사인유사도를 사용하여 화장품 성분의 텍스트 유사도를 파악해주고 유사도가 높은 순으로 정렬해주었다.

```
from sklearn.metrics.pairwise import cosine_similarity

text_sim=cosine_similarity(text_mat,text_mat)
text_sim_sorted_ind=text_sim.argsort()[::-1]
```

이후 사용자가 선택할 수 있는 옵션을 주어 상황에 맞는 상품을 추천해 주는 함수를 만들었다.

하나는 비건 옵션으로, 사용자가 비건 옵션을 선택한다면 식물성 화장품으로 분류된 화장품만을 추천해 주고, 비건 옵션을 선택하지 않는다면 동물성 화장품과 식물성 화장품을 구분 없이 추천해 준다.

다른 하나는 카테고리 옵션으로, 사용자가 카테고리 옵션을 선택한다면 입력한 화장품과 같은 카테고리의 상품만을 추천해 주고, 카테고리 옵션을 선택하지 않는다면 카테고리에 상관없이 비슷한 성분을 가진 화장품을 추천해 준다.

물론 두 가지 옵션을 동시에도 적용 가능하도록 하였다.

```
def find_sim_text(df,sorted_ind,product_name,top_n=10,vegan=0,cate=0):

    title_product=df[df['product']==product_name]
    title_index=title_product.index.values

    similar_indexes=sorted_ind[title_index]
    similar_indexes=similar_indexes[similar_indexes!=title_index]

    similar_indexes=similar_indexes.reshape(-1)
    indexes=similar_indexes

    # 비건 옵션
    if vegan==1:
        vegan_indexes=df.loc[df['animal']==0].index.values

    # 카테고리 옵션
    if cate==1:
        category=df.loc[df['product']==product_name,'cate'].values
        cate_indexes=df.loc[df['cate']==category[0]].index.values

    # 옵션 적용
    option_lst=[]
    if vegan==1 and cate==1: # 비건, 카테고리
        for idx in similar_indexes:
            if idx in vegan_indexes and idx in cate_indexes:
                option_lst.append(idx)
        indexes=option_lst

    elif vegan==1 and cate==0: # 비건
        for idx in similar_indexes:
            if idx in vegan_indexes:
```

```
option_lst.append(idx)
indexes=option_lst

elif vegan==0 and cate==1: # 카테고리
    for idx in similar_indexes:
        if idx in cate_indexes:
            option_lst.append(idx)
            indexes=option_lst

return df.iloc[indexes][:top_n]
```

작동 예시를 살펴보면 다음과 같다.

라운드어라운드의 수분 클렌징 오일을 추천시스템에 입력을 해보겠다.

	brand	product	text	cate	animal
1423	라운드어라운드	라운드어라운드 그린티 수분 클렌징 오일 300ml	호호바씨오일 오렌지껍질오일 카프릴릭 올리브오일 리모넨 리날롤 에틸헥실글리세린 시트르...	스킨케어	0

올리브영 추천시스템 예시 입력 상품

아무 옵션도 선택하지 않고 추천시스템을 돌린 결과는 다음과 같다.

	brand	product	text	cate	animal
1412	마녀공장	마녀공장 퓨어 클렌징 오일 200ml	호호바씨오일 들꽃오일 아이소아밀라우레이트 라벤더오일 카프릴릭 올리브오일 쥬빌레오일과물...	스킨케어	0
1420	스킨1004	스킨1004 마다가스카르 샌들라 라이트 클렌징 오일 200ml	올리브오일 호호바씨오일 리모넨 해바라기씨오일 에틸헥실글리세린 샌티드제라늄꽃오일 리날...	스킨케어	0
1395	에이트루	에이트루 퓨어 밸런싱 클렌징 오일 150ml 기획	올리브오일 호호바씨오일 트로폴론 리모넨 에틸헥실글리세린 리날롤 포도씨오일 베르가모트...	스킨케어	0
1391	에이트루	에이트루 퓨어 밸런싱 클렌징 오일 300ml	올리브오일 호호바씨오일 트로폴론 리모넨 에틸헥실글리세린 리날롤 포도씨오일 베르가모트...	스킨케어	0
2419	포맨트	포맨트 시그니처 퍼퓸 50ml (코튼블라썬부케)	제라니움 부틸렌글라이콜 리날롤 정제수 에탄올 시트로넬라 엑스신남알 향료 리모넨	향수	0
34	큐어	김정문알로에 큐어 리알로에 시그니처 크림 55g	솔비탄세스퀴올리에이트 들꽃오일 하이드로제네이티드폴리아이소부틴 알로에베라일즙 녹차씨오...	스킨케어	1
1353	더파이에센셜	더파이 선폴라워 클렌징 오일 200ml	아이스프로필미리스테이트 로즈마리오일 카프릴릭글라이콜 스쿠알란 카프릭트라이글리세라이...	스킨케어	0
2974	미장센	미장센 퍼팩트 노 워시 트리트먼트 미스트 250ml 대용량	토코페롤 하이드록시시트로넬알 글라이코리피드 호호바씨오일 동백나무씨오일 제라니움 연성...	헤어	0
2956	미장센	미장센 퍼팩트 오리지널 세럼 200ml	토코페롤 베타카로틴 글라이코리피드 호호바씨오일 동백나무씨오일 고추추출물 제라니움 카...	헤어	0
937	이즈엔트리	이즈엔트리 하이루존산 워터 미스트	베타글루칸 솔비탄세스퀴올리에이트 안하이드로자일러들 카프릴릭 리모넨 리날롤 팜틸렌글라...	스킨케어	0

아무 옵션 선택 안 한 경우

주로 입력한 상품과 같은 종류의 클렌징 오일을 많이 추천해 주었고, 그렇지 않은 상품에 대해서는 입력한 상품의 성분들과 비슷한 성분들
로 구성된 화장품을 추천해 주는 것을 확인할 수 있다.

카테고리 옵션을 선택한 결과는 다음과 같다.

	brand	product	text	cate	animal
1412	마녀공장	마녀공장 퓨어 클렌징 오일 200ml	호호바씨오일 들꽃오일 아이소아밀라우레이트 라벤더오일 카프릴릭 올리브오일 쥬빌레오일과물...	스킨케어	0
1420	스킨1004	스킨1004 마다가스카르 샌들라 라이트 클렌징 오일 200ml	올리브오일 호호바씨오일 리모넨 해바라기씨오일 에틸헥실글리세린 샌티드제라늄꽃오일 리날...	스킨케어	0
1395	에이트루	에이트루 퓨어 밸런싱 클렌징 오일 150ml 기획	올리브오일 호호바씨오일 트로폴론 리모넨 에틸헥실글리세린 리날롤 포도씨오일 베르가모트...	스킨케어	0
1391	에이트루	에이트루 퓨어 밸런싱 클렌징 오일 300ml	올리브오일 호호바씨오일 트로폴론 리모넨 에틸헥실글리세린 리날롤 포도씨오일 베르가모트...	스킨케어	0
34	큐어	김정문알로에 큐어 리알로에 시그니처 크림 55g	솔비탄세스퀴올리에이트 들꽃오일 하이드로제네이티드폴리아이소부틴 알로에베라일즙 녹차씨오...	스킨케어	1
1353	더파이에센셜	더파이 선폴라워 클렌징 오일 200ml	아이스프로필미리스테이트 로즈마리오일 카프릴릭글라이콜 스쿠알란 카프릭트라이글리세라이...	스킨케어	0
937	이즈엔트리	이즈엔트리 하이루존산 워터 미스트	베타글루칸 솔비탄세스퀴올리에이트 안하이드로자일러들 카프릴릭 리모넨 리날롤 팜틸렌글라...	스킨케어	0
57	스킨푸드	스킨푸드 블랙슈가 퍼팩트 에센셜 스크럽 2X 210g	호호바씨오일 마가부리추출물 스테아릴코능헥토라이트 카프릴릭 글리세린 파피아일메추출물...	스킨케어	1
1380	아비브	아비브 포어 클렌징 오일 어성초 오일 워시 200ml	폴리아이소부텐 호호바씨오일 라벤더꽃추출물 아이소프로필팔미테이트 클레라오일 카프릴릭...	스킨케어	0
1348	마녀공장	마녀공장 퓨어 클렌징 오일 200ml+퓨어폼 20ml	호호바씨오일 들꽃오일 아이소아밀라우레이트 라벤더오일 카프릴릭 올리브오일 쥬빌레오일과물...	스킨케어	0

카테고리 옵션 선택

카테고리 옵션을 선택하지 않았을 경우에는 향수와 헤어 카테고리에서도 상품이 추천된 것과 달리 카테고리 옵션을 선택한 경우 스킨케어
에서만 상품을 추천해 주는 것을 확인할 수 있다.

비건 옵션을 선택한 결과는 다음과 같다.

	brand	product	text	cate	animal
1412	마녀공장	마녀공장 퓨어 클렌징 오일 200ml	호호바씨오일 들궁오일 아이소아일라우레이트 라벤더오일 카프릴릭 올레브오일 썬발요아과물...	스킨케어	0
1420	스킨1004	스킨1004 마다가스카르 선텔라 라이트 클렌징 오일 200ml	올리브오일 호호바씨오일 리모넬 해바라기씨오일 에틸헥실글리세린 설탕드제라늄꽃오일 리날...	스킨케어	0
1395	에이트루	에이트루 퓨어 밸런싱 클렌징 오일 150ml 기획	올리브오일 호호바씨오일 트로폴론 리모넬 에틸헥실글리세린 리날롤 포도씨오일 베르가모트...	스킨케어	0
1391	에이트루	에이트루 퓨어 밸런싱 클렌징 오일 300ml	올리브오일 호호바씨오일 트로폴론 리모넬 에틸헥실글리세린 리날롤 포도씨오일 베르가모트...	스킨케어	0
2419	포맨트	포맨트 시그니처 퍼퓸 50ml (코튼딜라잇부케)	제라니올 부틸헥실글라이콜 리날롤 정제수 에탄올 시트로넬롤 헥실신남알 향료 리모넬	향수	0
1353	더피이에센셜	더피이 선텔라워 클렌징 오일 200ml	아이소프로필메리스테이트 로즈마리오일 카프릴릭글라이콜 스쿠알란 카프릴트라이글리세라이...	스킨케어	0
2974	미장센	미장센 퍼팩트 노 워시 트리트먼트 마스크 250ml 대용량	토코페롤 하이드록시시트로넬알 글라이코리피드 호호바씨오일 동백나무씨오일 제라니올 년션...	헤어	0
2956	미장센	미장센 퍼팩트 오리지널 세럼 200ml	토코페롤 베타카로틴 글라이코리피드 호호바씨오일 동백나무씨오일 고추추출물 제라니올 카...	헤어	0
937	이즈멘트리	이즈멘트리 허아루론산 워터 미스트	베타글루칸 솔비탄세스퀴올리에이트 안하이드로자일라톨 카프릴릭 리모넬 리날롤 편린헥글라...	스킨케어	0
2977	미장센	미장센 퍼팩트 오리지널 세럼 80ml	토코페롤 베타카로틴 글라이코리피드 호호바씨오일 동백나무씨오일 고추추출물 제라니올 카...	헤어	0

비건 옵션 선택

비건 옵션을 선택하지 않은 경우에는 동물성 성분이 포함된 화장품도 추천 리스트에 포함되어 있었지만, 비건 옵션을 선택한 경우에는 동물성 성분이 포함되지 않은 화장품만을 추천해 주는 것을 확인할 수 있다.

비건옵션과 카테고리옵션을 모두 선택한 결과는 다음과 같다.

	brand	product	text	cate	animal
1412	마녀공장	마녀공장 퓨어 클렌징 오일 200ml	호호바씨오일 들궁오일 아이소아일라우레이트 라벤더오일 카프릴릭 올레브오일 썬발요아과물...	스킨케어	0
1420	스킨1004	스킨1004 마다가스카르 선텔라 라이트 클렌징 오일 200ml	올리브오일 호호바씨오일 리모넬 해바라기씨오일 에틸헥실글리세린 설탕드제라늄꽃오일 리날...	스킨케어	0
1395	에이트루	에이트루 퓨어 밸런싱 클렌징 오일 150ml 기획	올리브오일 호호바씨오일 트로폴론 리모넬 에틸헥실글리세린 리날롤 포도씨오일 베르가모트...	스킨케어	0
1391	에이트루	에이트루 퓨어 밸런싱 클렌징 오일 300ml	올리브오일 호호바씨오일 트로폴론 리모넬 에틸헥실글리세린 리날롤 포도씨오일 베르가모트...	스킨케어	0
1353	더피이에센셜	더피이 선텔라워 클렌징 오일 200ml	아이소프로필메리스테이트 로즈마리오일 카프릴릭글라이콜 스쿠알란 카프릴트라이글리세라이...	스킨케어	0
937	이즈멘트리	이즈멘트리 허아루론산 워터 미스트	베타글루칸 솔비탄세스퀴올리에이트 안하이드로자일라톨 카프릴릭 리모넬 리날롤 편린헥글라...	스킨케어	0
1380	아비브	아비브 포어 클렌징 오일 여상초 오일 워시 200ml	올리바이소부틸 호호바씨오일 라벤더꽃추출물 아이소프로필알미테이트 글레리올 카프릴릭 ...	스킨케어	0
1348	마녀공장	마녀공장 퓨어 클렌징 오일 200ml+ 퓨어폼 20ml	호호바씨오일 들궁오일 아이소아일라우레이트 라벤더오일 카프릴릭 올레브오일 썬발요아과물...	스킨케어	0
1386	라운드랩	라운드랩 약공 클렌징오일 200ml (클렌저 20ml 증정)	시트릭에씨드 소듐아세티오네이트 스쿠알란 트라이소듐에틸헥다이아민다이석시네이트 썬가루 ...	스킨케어	0
921	브림그린	브림그린 달근 벤틱 참터 세럼 45ml	올리비질일수출물 알부민 잔탄검 베타버루리올 나이아신아마이드 오렌지껍질오일 글리세린...	스킨케어	0

비건옵션 카테고리 옵션 모두 선택

스킨케어만, 또한 동물성이 포함되지 않은 화장품들만을 추천해주는 것을 확인할 수 있다.

v. 알레르기 및 글루텐 포함 여부 파악

화장품을 사용하다 보면 피부에 직접적으로 닿는 제품이다 보니 성분뿐 아니라 알레르기에 대한 정보도 함께 찾아보게 된다. 또한 일반적으로 알려진 알레르기뿐 아니라 밀가루 알레르기도 등장하게 되면서 글루텐의 포함 여부도 파악하고자 하는 사람들이 증가하고 있다.

따라서 성분을 이용해서 분류하고 추천 시스템을 구현하는 과정에서 알레르기 및 글루텐 포함 여부를 함께 알려준다면 더 좋은 서비스가 될 것이라고 판단하여 이 기능을 추가하였다.

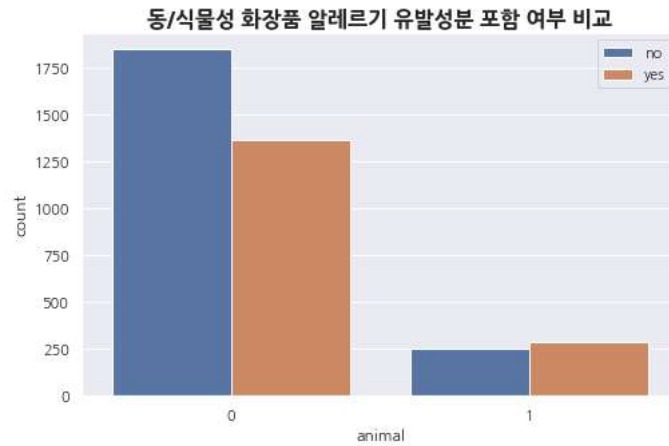
화장품에서 알레르기를 일으키는 종류에는 '아밀신남알', '벤질알코올', '신나밀알코올', '시트랄', '유제놀', '하이드록시시트로넬알', '이소유제놀', '아밀신나밀알코올', '벤질살리실레이트', '신남알', '쿠마린', '제라니올', '아니스에탄올', '벤질신나메이트', '파네솔', '부틸페닐메틸프로피오날', '리날롤', '벤질벤조에이트', '시트로넬롤', '헥실신남알', '리모넬', '메칠2옥티노에이트', '알파이소메칠이오논', '참나무이끼추출물', '나무이끼추출물' 등이 있다.

화장품 성분들 중 여기에 해당하는 성분이 존재한다면 알레르기 column에 따로 추가하였고, 알레르기 유발 성분이 존재하지 않는다면 '발견사항 없음'이라고 표시해 주었다.

```
def allergy_check(text):
    allergy_lst=[]
    for word in text.split(','):
        if word in cosmetic_allergy_lst:
            allergy_lst.append(word)
    if len(allergy_lst)>0:
        return ', '.join(allergy_lst)
    else:
        return '발견사항 없음'

olive['allergy']=olive['text'].apply(allergy_check)
```

다음 그래프는 동/식물성 화장품의 알레르기 유발 성분 포함 여부를 비교한 그래프이다.



올리브영 알레르기 데이터 시각화

식물성 화장품의 경우에는 알레르기 유발 성분을 포함한 상품보다 포함하지 않은 상품이 더 많이 존재했고, 동물성 화장품의 경우에는 알레르기 유발 성분을 포함하지 않은 상품보다 포함한 상품이 더 많음을 확인할 수 있다.

글루텐을 포함하고 있는 식품 및 성분에는 '글루테닌', '글리아딘', '보리', '귀리', '밀', '밀가루', '중력분', '강력분', '박력분' 등이 있다.

화장품 성분들 중 여기에 해당하는 성분이 존재한다면 글루텐 column에 따로 추가하였고, 글루텐이 포함되지 않은 상품이라면 '발견사항 없음'이라고 표시해 주었다.

```
def gluten_ckeck(text):
    gluten_lst=[]
    for word in text.split(','):
        if word in glu:
            gluten_lst.append(word)
    if len(gluten_lst)>0:
        return ','.join(gluten_lst)
    else:
        return '발견사항 없음'

olive['gluten']=olive['text'].apply(gluten_ckeck)
```

다음은 알레르기와 글루텐 포함 여부 파악까지 마친 올리브영 데이터 프레임이다.

	brand	product	text	cate	animal	allergy	gluten
0	렘시리즈	렘시리즈 울인원 트리트먼트 50ml 1+1 가릭세트	사이클로덱스트린,포타슘스테아레이트,포도씨추출물,스쿠알란,옥틸도데실 테오펜타노에이트,에...	스킨케어	1	발견사항 없음	발견사항 없음
1	유세린	유세린 허이얼루온 아이크림x나이트크림 기화(아이크림 15ml+나이트크림 50ml+건...	에실렉실리세린,부틸렌글리콜디카프릴레이트,이소프로필미티레이트,하이드로제네이티드코코리세...	스킨케어	1	발견사항 없음	발견사항 없음
2	아이디얼포맨	아이디얼 포맨 퍼펙트스킨케어 2종세트(마니어치 3종 중정)	제닐알라닌,무화과추출물,홀리베길알주출물,나이아신아마이드,카프릴릭,다이소다이티에어...	스킨케어	1	부틸테넨메틸프로파ionate,레오넨,리날롤	발견사항 없음
3	다슈	다슈 면즈 아쿠아 토너/로션 153ml 2종 세트(+토너&로션&글렌징 30ml중정)	개장각주출물,연꽃씨추출물,바실라스,마드리카리아꽃추출물,피루리발프추출물,제닐알라닌...	스킨케어	1	부틸테넨메틸프로파ionate,레오넨,리날롤	발견사항 없음
4	달바	달바 화이트 트러플 퍼스트 아로마틱 토너 155ml	테오펜텐글라이콜디미티라노에이트,하이드록시메틸우레아,등공오일,브로멜라인,소듐클로라이드...	스킨케어	1	리날롤,리오넨,핵살신날알	발견사항 없음
...
3748	대성디바	대성디바 메직프레스 베이스 실드	도코페론,스위트아몬드오일,아이소프로필알코올,해바라기씨오일,포스포리엑세이드,실리카,트라...	네일	0	발견사항 없음	발견사항 없음
3749	멧치유	멧치유 삼케어 합합 글로시&메드	유칼립투스잎추출물,하이드록시사이클로헥실페닐케톤,생물모오스세스테이트부티레이트,해마,메...	네일	0	발견사항 없음	발견사항 없음
3750	대성디바	대성디바 러지 세럼	벤질살리실레이트,포도씨오일,미네랄오일,스타이렌코폴리머,부틸렌,메틸렌,부틸테넨메틸프로...	네일	0	벤질살리실레이트,부틸테넨메틸프로파ionate,핵살신날알	발견사항 없음
3751	위드산	위드산 에코 네일 러무베 200ml	제라니올,리날롤,프로필렌글리콜,정제수,라벤다오일,도코페릴아세테이트,오레가노오일,로...	네일	0	제라니올,리날롤,리오넨	발견사항 없음
3752	위드산	위드산 메가사민애포드 15ml	메디칼에씨드,에틸아세테이트,트리클로릭만하이드라이드코폴리머,메탄올,나이트로글리세롤,포...	네일	0	발견사항 없음	발견사항 없음

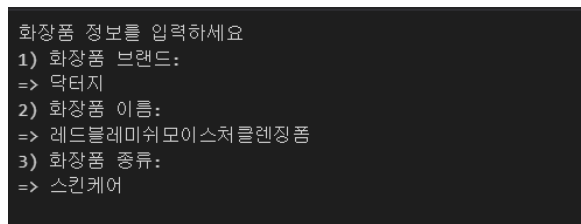
올리브영 데이터 알레르기, 글루텐 포함 여부

vi. 사용 예시

앞에서 만들어낸 코드들을 활용해서 사용자가 새로운 화장품을 입력했을 때 어떻게 진행되는지 그 흐름을 정리해 보았다.

1. 화장품 정보 및 사진 입력

먼저 사용자에게 새로운 화장품에 대한 정보를 입력 받는다.

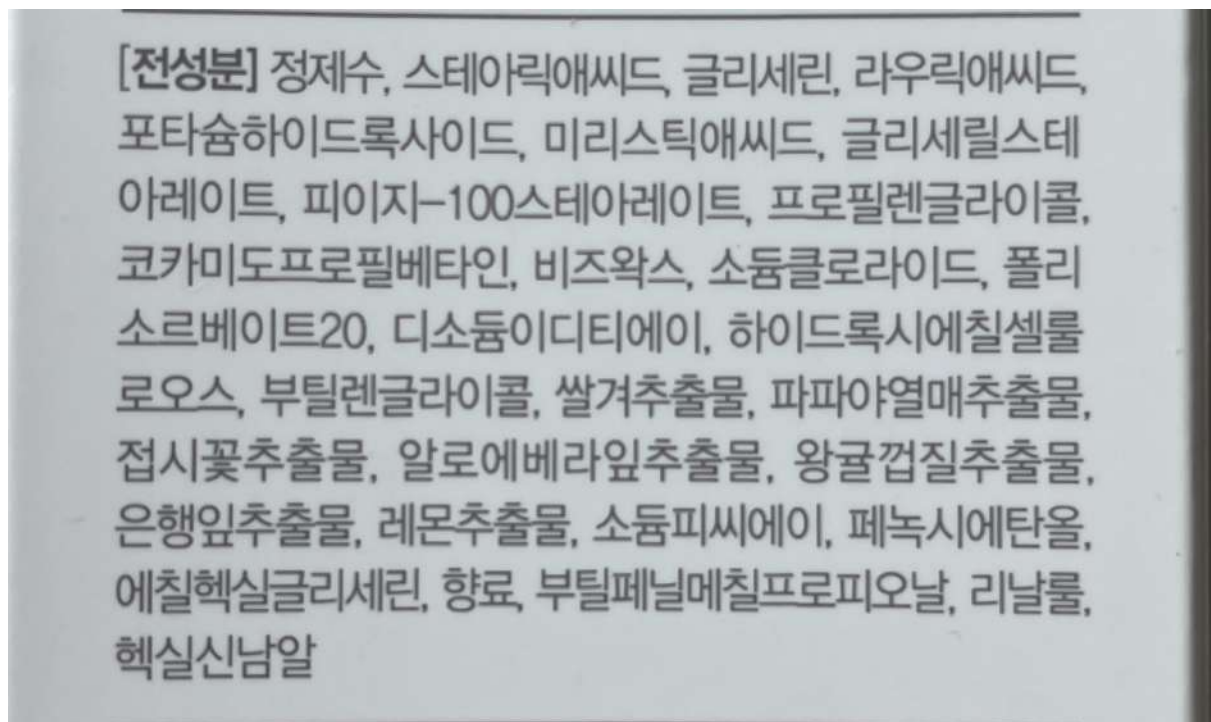


올리브영 사용에서 상품정보 입력



올리브영 사용에서 상품정보 입력 gif

화장품 성분은 사진을 찍어 입력받도록 한다. pytesseract를 통해 입력받은 이미지에서 텍스트를 추출하고 전처리를 거쳐 성분에 대한 데이터만 살려준다.



새로운 화장품 성분 사진

이렇게 입력받은 데이터들로 입력 데이터프레임을 완성한다.

brand	product	text	cate
0 닥터지	레드 블레쉬워 모이스처 클렌징 폼	,정제수,스테아릭애씨드,글리세린,라우릭애씨드,포타슘하이드록사이드,미리스틱애씨드,글리세릴스테아레이트,피이지-100스테아레이트,프로필렌글라이콜,코카미도프로필베타인,비즈왁스,소듐클로라이드,폴리소르베이트20,디소듐이디티에이,하이드록시에틸셀룰로오스,부틸렌글라이콜,쌀겨추출물,파파야열매추출물,접시꽃추출물,알로에베라잎추출물,왕굴껍질추출물,은행잎추출물,레몬추출물,소듐피씨에이,페녹시에탄올,에칠헥실글리세린,향료,부틸페닐메칠프로피오날,리날룰,헥실신남알	스킨케어

올리브영 사용에서 입력 상품 정보 데이터프레임

2. 동물성 / 식물성 화장품 분류

입력받은 성분들 중에 동물성 성분이 포함되었는지 확인한다. 동물성 성분이 포함되어 있다면 '동물성 성분이 포함되어 있어요.'라는 메시지를, 포함되지 않았다면 '동물성 성분이 발견되지 않았어요.'라는 메시지를 출력한다.

[레드 블레미쉬 모이스처 클렌징 폼]에는 동물성 성분이 포함되어 있어요.

올리브영 사용예시 동물성/식물성 분류 결과

3. 알레르기 확인

입력받은 성분들 중에 알레르기를 유발하는 성분이 포함되어 있는지 확인한다. 알레르기 유발 성분이 포함되어 있다면 포함된 알레르기 정보와 함께 알레르기 성분이 들어있다는 메시지를, 알레르기 유발 성분이 포함되어 있지 않다면 알레르기 성분이 발견되지 않았다는 메시지를 출력한다.

[레드 블레미쉬 모이스처 클렌징 폼]에는 알레르기 성분 ['리날롤']이 들어있어요.

올리브영 사용예시 알레르기 결과

4. 글루텐 확인

입력받은 성분들 중에 글루텐이 포함되어 있다면 '글루텐이 포함되어 있습니다.'라는 메시지를, 글루텐이 포함되어 있지 않다면 '글루텐이 포함되어 있지 않습니다.'라는 메시지를 출력한다.

[레드 블레미쉬 모이스처 클렌징 폼]에는 글루텐이 포함되어 있지 않습니다.

올리브영 사용예시 글루텐 결과

5. 추천시스템

입력 받은 성분들과 비슷한 성분을 가진 제품을 추천하기에 앞서 사용자가 원하는 옵션을 먼저 입력 받는다.

카테고리 옵션을 선택하세요
1) 같은 카테고리 상품만 궁금해요
2) 다른 카테고리 상품도 궁금해요
=> 1번 옵션 선택

비건 옵션을 선택하세요
1) 동물성 성분이 없는 제품만 궁금해요
2) 동물성 상관없이 모두 궁금해요
=> 1번 옵션 선택

올리브영 사용예시 추천시스템 옵션 입력

```
# 옵션 입력
print('카테고리 옵션을 선택하세요')
print('1) 같은 카테고리 상품만 궁금해요')
print('2) 다른 카테고리 상품도 궁금해요')
cate_option=input('1 / 2')
print('<-> 'cate_option'번 옵션 선택')

if cate_option=='1':
    cate_option=1
else:
    cate_option=2

print('비건 옵션을 선택하세요')
print('1) 동물성 성분이 없는 제품만 궁금해요')
print('2) 동물성 상관없이 모두 궁금해요')
vegan_option=input('1 / 2')

print('<-> 'vegan_option'번 옵션 선택')
if vegan_option=='1':
    vegan_option=1
else:
    vegan_option=2

recom_df=pd.concat([input_data,df]).reset_index(drop=True)
count_vector=countvectorizer(mle_df,vgan_range(1,2))
text_mat=count_vect.fit_transform(recom_df['text'])
text_sim=cosine_similarity(text_mat,text_mat)
text_sim_sorted_ind=text_sim.argsort()[::-1]

print('입력한 상품과 비슷한 제품 5가지를 보여드립니다')
find_sim_text(recom_df,text_sim_sorted_ind,input_data['product'].values[0],5,vegan_option,cate_opt
```

올리브영 사용예시 추천시스템 옵션 입력 gif

사용자의 선택 옵션에 맞는 추천 상품들을 출력해준다.

	brand	product	text	cate	animal	allergy	gluten
1325	삼촌	삼촌정고 내수원 글렌징용 다불 기원 (180ml+180ml)	포타슘하이드록사이드 레온그라스추출물 라우릭에씨드 페퍼민트잎추출물 글리세린 리모넨 리...	스킨케어	0	리모넨,리날룰	발견사항 없음
1290	라로슈포제	라로슈포제 풀러리앙 퓨리피잉 포밍 크림	미리스틱에씨드 테트라소듐이디티에이 판미틱에씨드 포타슘하이드록사이드 정제수 라우릭에씨...	스킨케어	0	발견사항 없음	발견사항 없음
1264	라로슈포제	라로슈포제 풀러리앙 포밍글렌저 다불 기원	미리스틱에씨드 테트라소듐이디티에이 판미틱에씨드 포타슘하이드록사이드 정제수 라우릭에씨...	스킨케어	0	발견사항 없음	발견사항 없음
1276	스트라이엑스	스트라이엑스 약알칼리성 비하 폼클렌저 기원 (150ml+15ml)	포타슘하이드록사이드 소듐글로라이드 무화과추출물 미역추출물 예뵈멘틴카복사마이드 라우릭...	스킨케어	0	리모넨,리날룰	발견사항 없음
1262	아크네스	아크네스 클리어&화이트 포밍워시 1+1 기원	시트릭에씨드 판미틱에씨드 포타슘하이드록사이드 라우릭에씨드 에탄올 글리세린 다이아포타슘...	스킨케어	0	리모넨,리날룰	발견사항 없음
1279	한울	한울 어민썬 진정 맑은 클렌징폼 기원세드 (폼120g + 수문진정크림15g)	팔미틱에씨드 포타슘하이드록사이드 라우릭에씨드 글리세릴스테아레이트 글리세린 다이아소듐...	스킨케어	0	발견사항 없음	발견사항 없음
1307	라로슈포제	라로슈포제 풀러리앙 포밍글렌저 125ml 기원	미리스틱에씨드 테트라소듐이디티에이 증정풀러리앙 판미틱에씨드 포타슘하이드록사이드 정제...	스킨케어	0	발견사항 없음	발견사항 없음
1336	스킨푸드	스킨푸드 블랙슈가 파렉트 베킵폼 200ml	마키부러추출물 라우릴포스페이트 포타슘하이드록사이드 소듐글로라이드 라우릭에씨드 글리세...	스킨케어	0	리모넨,시드랄	발견사항 없음
1273	브림그린	브림그린 티트리시가 트러블클렌징폼 더블기원 (200ml+200ml)	베타글루칸 판미틱에씨드 포타슘하이드록사이드 라우릭에씨드 글리세릴스테아레이트 글리세린...	스킨케어	0	발견사항 없음	발견사항 없음
1304	뉴트로지나	뉴트로지나 딥글린 쉼폼 포밍 클렌저 220g+50g 기원	다이아소듐이디티에이 미리스틱에씨드 소듐코코알글라이시네이트 코카메도프로필베타딘 정제수 ...	스킨케어	0	발견사항 없음	발견사항 없음

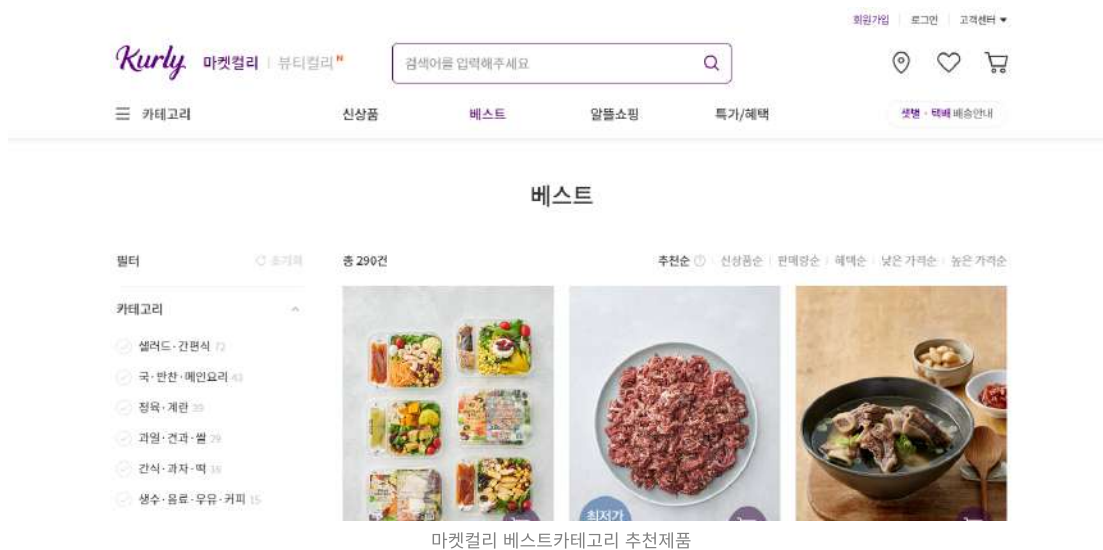
올리브영 사용에서 추천시스템 결과

c. 비건 식품 및 화장품 후기 모음집

i. 마켓컬리

1. 데이터 크롤링

성분표 분석을 통해 구매하려는 채식 식품의 리뷰까지 한번에 볼 수 있는 서비스를 제공하여 사용자의 편의를 도모하기 위해 마켓컬리 베스트 제품 100가지의 리뷰 100개를 크롤링 작업을 하였다. 다양한 제품 분석을 위해 비건에 국한시키지 않고 베스트 제품을 가져오기로 하였다.



마켓컬리 베스트카테고리 추천제품

상품설명	상세정보	후기 (164,783)	문의
공식 김주혁 메이크업 후기 안내			
공식 상품 후기 적립금 정책 안내			
<div> <div> <div>베스트</div> <div>알만 김**</div> </div> <div> <div>[스윗밸런스] 닭가슴살 고구마 샐러드</div> <div> <p>가격도 저렴하고 중량에 비해 칼로리도 적당한 편이라 너무 좋아요!</p> <p>하루 샐러드 하나는 꼭 먹고 있는데 여러가지 맛 사봤는데 스윗밸런스 닭가슴살 고구마 샐러드 너무 맛있네요!</p> <p>닭가슴살은 엄청 촉촉하고 약간 짭조름하게 간이 되어 있어요.</p> <p>저는 빨간 토마토 보다 다른색 토마토를 더 좋아하는데 기분탓일지 모르겠지만 버린맛이 덜 하다고 해야할까? 그래서 더 좋았어요!</p> <p>매주리얼이 물릴 때쯤 새로운 맛으로 입안에 씹혀지는게 너무 맛있고 매력적 이었는데요.</p> <p>병아리콩이랑 샐러드가 적절히 섞여 있어서 너무 맛있었어요!</p> <p>이거 하나 먹었는데 포만감이 아주 상당해요 든든합니다. 좀 부족하신 분들은 채소를 따로 추가해서 드셔도 될 것 같아요.</p> <p>시저 드레싱 소스가 은근 양이 많고 맛도 있어서 아주 굿이거든요! 스윗밸런스 기억해줬다가 자주 시켜먹어야 겠어요!</p> </div> <div>  <div>상품 선택 ^</div> </div> </div> <div>제품 페이지 내의 후기카테고리</div> </div>			

베스트 카테고리 페이지 url을 받아온 다음 제품 이름과 리뷰를 크롤링 해오고 다음페이지로 반복하는 코드를 작성하였다.
받아와지는 페이지 갯수도 확인해가며 크롤링하였다.

```

from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver import ActionChains
from selenium.webdriver.common.by import By

import re
import pandas as pd
import time
import requests

url='https://www.kurly.com/collections/market-best'
driver=webdriver.Chrome()
driver.get(url)
act=ActionChains(driver)

product=[]
review_lst=[]
for i in range(1,97):
    driver.find_elements(By.CSS_SELECTOR, '#container > div > div.css-1d3w5wq.ef36txc6 > div.css-11kh0cw.ef36txc5 > a:nth-child('+str(i)
    time.sleep(1)

    # 상품 이름
    html=driver.page_source
    soup=BeautifulSoup(html, 'html.parser')
    product.append(soup.select('div>h1.css-1f2zq3n.ezpe9l11')[0].text)
    review=[]

    # 리뷰
    try:
        driver.find_elements(By.CSS_SELECTOR, '#top > div.css-n48rgu.ex9g73v0 > div.css-16c0d8l.e1brqtzw0 > nav > ul > li:nth-child(3)'
        while(1):
            for idx in range(10):
                html=driver.page_source
                soup=BeautifulSoup(html, 'html.parser')
                review.append(soup.select('p.css-i69j0n.e36z05c5')[idx].text)
            # 다음페이지
            driver.find_elements(By.CSS_SELECTOR, '#review > section > div.css-1nrf0nk.e1kog1is3 > div.css-jz9m4p.e1kog1is5 > button.c
            time.sleep(1)
            if len(review)>100:
                break
    except:
        review.append('')
    review_lst.append(review)

    # 메인 화면으로
    driver.back()

    # 진행상황 체크
    if i%10==0:

```

```
print(i)

df=pd.DataFrame({'product':product,'review':review_lst})
df
```

위의 코드를 통해 아래의 데이터 프레임이 출력되었다.

product 컬럼에 제품이름이, review에는 한 제품당 100개의 리뷰가 들어갔다. 크롤링 되지 않은 제품과 리뷰를 제외하고 총 96개의 제품 10500개의 리뷰로 구성되었다.

	product	review
0	[사미현] 갈비탕	[사미현 갈비탕 추천이용!!\n침 먹어봤는데.. 재구매각 바로 섰습니다.:ㅋㅋ\n사...
1	[KF365] 양념 소불고기 1kg (냉장)	[처음 배송 받았을 땐 1키로가 생각보다 적네? 했는데.. 웬걸.. 팬에 다 볶고 ...
2	[이연복의 목란] 짬뽕 2인분 (맵기선택)	[마켓컬리에서 제일 자주 주문한 냉동식품 중 1위가 목란 짬뽕일 정도로 정말 좋아해...
3	[태우한우] 1+ 한우 안심 구이용 200g (냉장)	[안심 스테이크는 늘 태우한우에서 사먹어요 일단 투블은 지방함량이 높아 건강면에서도...
4	[Kurly's] 동물복지 유정란 20구	[\n\n 샷배배송이 아닌 택배배송 지역인데다가 인터넷으로 계란을 주문해 본 적이...

마켓컬리 리뷰 크롤링 데이터프레임

2. 데이터 전처리

한 행에 100개의 리뷰가 들어있어 어떻게 꺼내서 전처리할 지에 대한 다양한 시도와 실패가 있었다. stack을 이용해 모든 리뷰를 행으로 만들어 전처리 하는 과정은 코드가 지저분하게 보이는 단점이 있었다. 그 결과 가장 마지막에 작성한 아래의 코드를 통해 전처리 하였다.

review를 한 행씩 가져온 다음 정규표현식을 사용하여 띄어쓰기 될 때 사용된 \n을 삭제하고 영어와 숫자 특수문자는 제외하였다.

다만 대괄호까지 삭제될 경우 리스트문에서 벗어나기 때문에 대괄호는 포함시키도록 하였다.

```
from pprint import pprint
import re

for i in range(96):
    product_lst = df['product'][i]

    for lst in product_lst:
        lst = []

    #정규표현식
    for i in range(96):
        review = df['review'][i]
        review =review.replace('\n','')
        review =review.replace('n','')
        review = review.replace(' ',' ')
        review = re.sub('[^가-힣.,\[\]]',' ', str(review))
        review = review.strip()
        lst.append(review)

df["review"] = lst
```

전체 리뷰 리스트 안에 각 제품별 리스트를 넣었다. 함께 들어있기 때문에 strip등 공백 제거가 잘 작동하지 않는 점이 아쉬움으로 남는다.

또한 맞춤법 검사기 모듈 hanspell과 띄어쓰기 모듈 pyokospacing이 지속적인 오류로 작동하지 않아 적용시키지 못하였다.

심각한 오타는 없는 편이기에 다음 과정으로 넘어가는데 문제가 생기진 않았다.

	product	review
0	[사미한] 갈비탕	[사미한 갈비탕 추천이용 첨 먹어봤는데.. 재구매각 바로 섰습니다. 사실 저는 ...
1	[KF365] 양념 소불고기 1kg (냉장)	[처음 배송 받았을 땐 키로가 생각보다 적네 했는데.. 웬걸.. 팬에 다 볶고...
2	[이연복의 묵란] 짬뽕 2인분 (맬기선택)	[마켓컬리에서 제일 자주 주문한 냉동식품 중 위가 묵란 짬뽕일 정도로 정말 좋아...
3	[태우한우] 1+ 한우 안심 구이용 200g (냉장)	[안심 스테이크는 늘 태우한우에서 사먹어요 일단 투블은 지방함량이 높아 건강면에서...
4	[Kurly's] 동물복지 유정란 20구	[셋별배송이 아닌 택배배송 지역인데다가 인터넷으로 계란을 주문해 본 적이 ...
...
91	[최현석의 초이닷] 새우 봉골레 파스타	[늦잠 실패 자고 일어 나서 아침에 냉장고로 옮겨 둔 초이닷 봉골레 파스타를 꺼내...
92	[그릭데이] 그릭요거트 시그니처 450g	[꾸덕꾸덕한 그릭요거트 너무 좋아해서 몇번째 사먹는지 몰라요. 마켓컬리가 신선한 ...
93	[연안식당] 부추 고막 비빔장	[폭죽 찢는 폭음에 저녁밥 차리는게 제일 힘든 요즘 나가서 먹기도 귀찮고 연안...
94	냉동 블루베리 1kg (철레산)	[요즘 커피 덕에 냉동실에 빈틈이 없어용 왜 이렇게 맛있는걸 많이 파시는지 오히...
95	[속초해품] 백명란젓 400g	[명란젓 하나 사서 두루두루 잘 활용하고 있어요.좋은 점이 저염이라 많이 넣어도 ...

96 rows × 2 columns

리뷰 전처리 후 데이터프레임

불용어 처리를 위해 한글 불용어 리스트 url을 가져와서 적용시킨 후 append를 통해 출력되지 않기를 원하는 글자를 더 추가해주었다.

```
stopwords = pd.read_csv("https://raw.githubusercontent.com/yoontk200/FastCampusDataset/master/korean_stopwords.txt").values.tolist()
review_stopwords = ['입니다', '있어요', '유정', '묵란', '달걀', '계란', '먹기', '컬리', '사과', '같아요']
for word in review_stopwords:
    stopwords.append([word])
```

konlpy의 okt 형태소 분석기를 이용하여 리뷰를 명사와 형용사만 출력되도록 하였다.

한 글자만 출력 되는 경우 의미 있는 데이터 일 가능성이 낮기 때문에 한 글자만 나올 경우 출력 되지 않게 작성하였고 가장 많이 나오는 단어 100개를 저장하고 순서대로 정렬하도록 하였다.

한글 형태소를 제일 잘 분리해준다는 mecab은 작동이 되지 않아 사용하지 못하였다.

```
from konlpy.tag import Okt
from wordcloud import WordCloud
from collections import Counter
import pandas as pd
text = df

for i in range(96):
    msg = text['review'][i]

    # # Okt 함수를 이용해 형태소 분석
    okt = Okt()
    line = []

    line = okt.pos(msg)

    n_adj = []
    # 명사 또는 형용사인 단어만 n_adj에 넣어주기
    for word, tag in line:
        if tag in ['Noun', 'Adjective']:
            n_adj.append(word)
    # 명사 또는 형용사인 단어 및 2글자 이상인 단어 선택 시
    n_adj = [word for word, tag in line if tag in ['Noun', 'Adjective'] and len(word) > 1]

    #불용어
    stop_words = stopwords

    n_adj = [word for word in n_adj if not word in stop_words]

    # 가장 많이 나온 단어 100개 저장
    counts = Counter(n_adj)
    tags = counts.most_common(100)

    print(tags)
```


0 [사미한 갈비탕 추천이용 첨 먹었는데... 재구매하 바로 썼습니다. 사실 저는 사미한 이 부산에 유명한 고깃집인것도 몰랐고, 살짝 세일하길래 한번 먹어 본진
1 [[갈비탕', 44), ('고기', 21), ('맛있어요', 21), ('국물', 16), ('아이', 14), ('구매', 14), ('주문', 14), ('미친', 13), ('좋아요', 12), ('자주', 11)
2 [처음 배송 받았을 땐 키가 생각보다 적네 했는데... 웰즈... 판매 다 못가 엄청 많은 양이더러구요... 고기가 풍족 있게는 분...미미 고기도 많지만 존
3 [[불고기', 34), ('맛있어요', 20), ('좋아요', 16), ('구매', 15), ('자주', 12), ('분해', 12), ('야채', 11), ('양도', 11), ('아이', 10), ('고기', 9)
4 [마켓컬리에서 제일 자주 주문한 냉동식품 중 위가 목란 팜필 청로 정팔 좋아해요. 거의 종주자 수준으로 냉동실에서 꼭 찾아들니다. 목란 팜품 국물은 맑고
5 [[참주', 54), ('국물', 33), ('맛있어요', 27), ('육란', 13), ('영도', 13), ('매름', 12), ('좋아요', 12), ('주문', 11), ('매운', 11), ('추천', 10)
6 [안심 스테이크는 늘 태우왕우에서 사먹어요 일단 토폴은 지방함량이 높아 건강면에서도 나쁘고... 그렇다고 아랫등골을 먹거나 입에서 맛있는 것도 중요하다 싶
7 [[안심', 23), ('맛있어요', 22), ('한우', 18), ('부드럽고', 17), ('구매', 15), ('맛있게', 15), ('좋아요', 15), ('고기', 9), ('자주', 9), ('아이', 9)
8 [쇠불배송이 아닌 태백배송이 지역인데다가 인터넷으로 제란을 주문해 본 적이 없었어요. 대형마트물에서 주문하러간 해 봤는데, 그전 바로 당일날 배송으
9 [[계란', 53), ('구매', 25), ('동물복지', 22), ('유정', 21), ('좋아요', 21), ('달걀', 19), ('노른자', 17), ('배출', 11), ('컬리', 10), ('항상', 10)
10 [생각했던것보단 양이 적었어요근데 너무 맛있어서 허겁지겁 먹었어요마켓컬리 처음 가입했는데 베스트에 쭈꾸미가 있더라구요남편이랑 퇴근하고 티비보면서 반죽
11 [[쭈꾸미', 30), ('맛있어요', 22), ('좋아요', 20), ('주문', 15), ('구매', 14), ('야채', 14), ('먹기', 12), ('맛있어서', 11), ('양념', 11), ('매콤', 11)
12 [치마살인데 호추산이든 원래 짜지 않은 부위인데, 까도 칼같이 해서 꽤나 칼리직인 가격에 구매했습니다. 그래서 큰 기대는 없었는데... 육질도 괜찮고 지
13 [[맛있어요', 21), ('고기', 16), ('구매', 14), ('남새', 13), ('좋아요', 10), ('자주', 10), ('부드럽고', 9), ('맛있게', 8), ('아이', 8), ('만나', 8)
14 [가격도 저렴하고 중량에 비해 칼로리도 적당한 편이라 너무 좋았어요 하루 셀러드 하나는 꼭 먹고 있는데 여러가지 맛 시켰는데스윗벨런스 닭가슴살 고구마 샐
15 [[샐러드', 53), ('맛있어요', 43), ('스틱', 31), ('이여', 31), ('여러가지', 29), ('구매', 27), ('가장', 27), ('좋아요', 14), ('먹기', 13), ('입리', 13)
16 [총합니다. 자주구입해서 먹어요 , 육합한 한데 약간 심심한 느낌이라 재구매는 많지 고민해볼 것 같아요. 간단한 아침이나 간식으로 좋아요. , 아침마다
17 [[좋아요', 24), ('맛있어요', 20), ('먹기', 20), ('아침', 19), ('치즈', 16), ('아이', 13), ('하나', 8), ('간식', 7), ('대용', 7), ('간단함', 6), ('
18 [트라이스 딸기 공주 채영이 멤버들에게 만들어 쪼든 딸기 산타 요리 레시피를 편들을 위해 알려줬어요 딸기의 상큼 달콤함 치즈의 담백함 하몽의 짭조름함
19 [[딸기', 58), ('맛있어요', 14), ('좋아요', 10), ('구매', 8), ('가크', 8), ('산타', 7), ('상대', 7), ('치즈', 6), ('맛있게', 6), ('선선하고', 6)
20 [딸기 공주 채영이 멤버들에게 만들어 쪼든 딸기 산타 요리 레시피를 편들을 위해 알려줬어요 딸기의 상큼 달콤함 치즈의 담백함 하몽의 짭조름함

3. 데이터 시각화

어떤 제품의 워드클라우드인지 알아보기 위해 제품이름과 종류는 불용어처리 하지 않았다.

```
import numpy as np
from PIL import Image

# # # WordCloud(워드클라우드) 만들기###
# # 폰트지정
font='C:/windows/Fonts/nanumgothicextrabold.ttf'
word_cloud = WordCloud(font_path ='malgun.ttf', background_color='white',max_font_size=400,width=800, height=600).generate_from_frequencies(word_frequencies)

# # 워드클라우드 사진으로 저장
word_cloud.to_file(f"wordcloud_images/wordcloud_{i}.png")

# # 사이즈 설정 및 화면에 출력
import matplotlib.pyplot as plt
plt.figure(figsize=(10,10))
plt.imshow(word_cloud)
plt.axis('off')
plt.show()
```

Vegan? 너도 할 수 있어



ii. 올리브영

```
from bs4 import BeautifulSoup
from selenium import webdriver
from selenium.webdriver import ActionChains
from selenium.webdriver.common.by import By

import re
import pandas as pd
import time
import requests

brand=[]
product=[]
score_lst=[]
review_lst=[]
for line in range(1,26):
    for i in range(4):
        url='https://www.oliveyoung.co.kr/store/main/getBestList.do'
        driver=webdriver.Chrome()
        driver.get(url)
        act=ActionChains(driver)
        html=driver.page_source
        soup=BeautifulSoup(html, 'html.parser')

        # 상세페이지 진입
        driver.find_elements(By.CSS_SELECTOR, '#Container > div.best-area > div.TabsConts.on > ul:nth-child('+str(line)+') > li')[i].click()

        # 브랜드, 상품 이름
        html=driver.page_source
        soup=BeautifulSoup(html, 'html.parser')
        brand.append(soup.select('div.prd_info > p > a')[0].text)
        product.append(soup.select('div.prd_info > p.prd_name')[0].text)
        review=[]
        score=[]

    try:
        # 리뷰창 이동
        driver.find_elements(By.CSS_SELECTOR, '#reviewInfo')[0].click()
        time.sleep(2)

        # 최신순으로 보기
        driver.find_elements(By.CSS_SELECTOR, '#gdasSort > li:nth-child(3)')[0].click()
        time.sleep(1)

        # 리뷰 크롤링
        for idx in range(10):
            html=driver.page_source
            soup=BeautifulSoup(html, 'html.parser')
            time.sleep(1)
            score.append(soup.select('span.review_point')[idx+1].text)
```

이러한 크롤링 과정을 통해 수집한 리뷰 데이터를 마켓컬리와 같은 과정으로 정제하여 워드클라우드를 만들었다.




마히나 비건 테이블 Mahina Ve...
 4.2 ★★★★★ (43) · 채식(Vegan)
 23.5 km · 서울특별시 강남구
 영업 종료 · 오전 11:30에 영업 시작
 매장 내 식사 · 테이크아웃 · 배달 서비스



구글맵 비건식당 검색 결과 예시

리베르테 Liberte
 4.4 ★★★★★ (153) · 프랑스 요리
 24.8 km · 서울특별시 강남구
 영업 종료 · 오후 12:00에 영업 시작
 "(페스코 베지테리언이라) 예약때 미리 말하면 해산물 요리를 준비해준다 ..."



구글맵 페스코 식당 검색 결과 예시

구글맵을 통해 채식 혹은 페스코 식단이 가능한 식당들을 크롤링 하여 사용자들이 비건 식단 가능 식당을 한눈에 보기 쉽게 지도에 시각화 하기로 하였다.

정렬 기준 관련도순 ▾

러브 더 비건즈
 리뷰 없음 · 채식
 1.0 km · 호명동 642-2
 매장 내 식사 · 테이크아웃

마히나 비건 테이블 Mahina Ve...
 4.2 ★★★★★ (43) · 채식(Vegan)
 23.5 km · 서울특별시 강남구
 영업 종료 · 오전 11:30에 영업 시작
 매장 내 식사 · 테이크아웃 · 배달 서비스

초록뜰
 4.2 ★★★★★ (109) · ₩ · 채식(Vegan)
 16.9 km · 서울특별시 동대문구
 영업 종료 · 일 오전 11:00에 영업 시작
 매장 내 식사 · 테이크아웃 · 배달이 안 됨

밥풀꽃
 리뷰 없음 · 채식(Vegan)
 16.3 km · 상계동 966-1번지 1층 송현빌...
 매장 내 식사 · 테이크아웃

베제투스
 4.4 ★★★★★ (224) · ₩₩₩ · 채식(Vegan)
 25.0 km · 서울특별시 용산구
 영업 종료 · 오후 12:00에 영업 시작
 매장 내 식사 · 테이크아웃 · 배달 서비스

더 많은 장소 보기

마히나 비건 테이블 Mahina Vegan Table ✕
 4.2 ★★★★★ (43) · 비건 채식 레스토랑



사진 120장 이상

경로 저장 전화 걸기

개요 메뉴 리뷰

서비스 옵션: 매장 내 식사 · 테이크아웃 · 배달 서비스

주소: 서울특별시 강남구 논현로175길752층

영업시간: 영업 종료 · 오전 11:30에 영업 시작 ▾

메뉴: mahinavegan.com

예약: naver.com 제공업체

주문: naver.com 제공업체

연락처: 050-71371-5331

수정 제안하기 · 이 비즈니스의 소유주인가요?

누락된 정보 추가

웹사이트 추가

구글맵 비건 식당 크롤링 화면

구글맵에 검색을 하면 목록에 식당을 20개씩 보여준다. 그 목록에서 식당 이름과 별점, 식당의 종류를 크롤링 한 후 식당을 클릭하여 상세 페이지를 열어주었다. 상세페이지에서는 자세한 주소를 크롤링 해오는 코드를 작성하였다.



이후 다음 버튼을 클릭하여 다음 목록을 불러와 크롤링 하는 작업을 다음 버튼이 더 이상 등장하지 않을 때까지 반복하여 주었다.

```
star=[]
name=[]
address=[]
try:
    while(1):
        for i in range(0,20):
            # 가게이름, 별점
            html=driver.page_source
            soup=BeautifulSoup(html, 'html.parser')
            for idx, info in enumerate(soup.select('div.rllt__details')[i]):
                if idx==0:
                    name.append(info.text)
                elif idx==1:
                    star.append(info.text)

            # 상세페이지 열기
            driver.find_elements(By.CSS_SELECTOR, 'div.uMdZh.tIXNaf>div>div>a>div>div')[i].click()
            time.sleep(1)

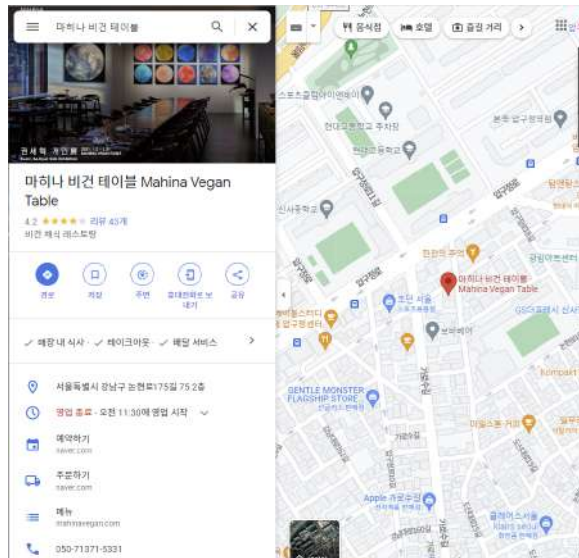
            # 주소
            html=driver.page_source
            soup=BeautifulSoup(html, 'html.parser')
            address.append(soup.select('span.LrzXr')[0].text)

            # 페이지 이동
            driver.find_elements(By.CSS_SELECTOR, '#pnnext > span:nth-child(2)')[0].click()
            time.sleep(1)

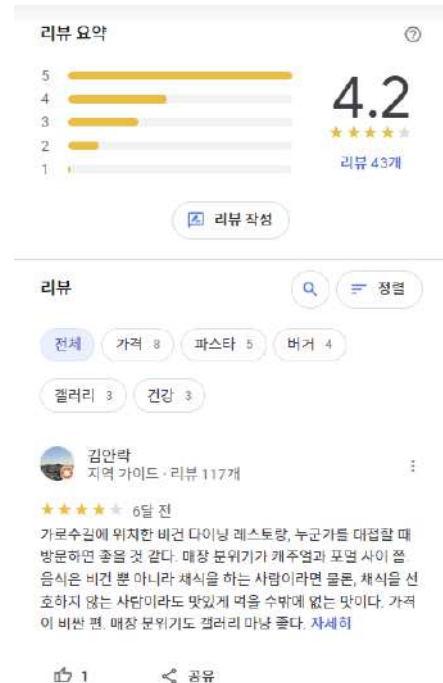
            # break문
            if len(name)>1000:
                break

            # 진행상황
            if len(name)%20==0:
                print(len(name))
except:
    print('end')
```

이후 구글맵에 식당 이름을 다시 검색해 보며 식당에 대한 더 자세한 정보와 리뷰를 얻을 수 있는 페이지의 url을 얻어오는 작업도 같이 진행하였다.



구글맵 식당 자세한 정보



```
url_lst=[]
for name in name_lst:
    url='https://www.google.com/maps/place/'+name
    driver=webdriver.Chrome()
    driver.get(url)
    act=ActionChains(driver)
    time.sleep(3)
    driver.find_elements(By.CSS_SELECTOR, '#searchbox-searchbutton')[0].click()
    time.sleep(3)
    url_lst.append(driver.current_url)
    driver.close()
```

같은 방식으로 구글맵에 '페스코식당'이라고 검색했을 때 나오는 식당들도 크롤링 하여 비건 식단 가능 데이터 프레임을 완성하였다.

	name	address	star	vegan	url	pesco
0	리브 더 베건즈	경기도 남양주시 호평동 642-2	리뷰 없음 · 채식	1	https://www.google.com/maps/place/%EB%9F%AC%EB...	NaN
1	망량비행국수	경기도 남양주시 호평동 농울2로14번길 29	4.2(20) · 음식점	1	https://www.google.com/maps/search/%EB%A7%9D%E...	NaN
2	대히나 비건 테이블 Mahina Vegan Table	서울특별시 강남구 논현로175길 75 2층	4.2(43) · 채식(Vegan)	1	https://www.google.com/maps/place/%EB%A7%88%ED...	NaN
3	밥돌곳	상계동 966-1번지 1층 송천별당 노원구 서울특별시 KR	리뷰 없음 · 채식(Vegan)	1	https://www.google.com/maps/search/%EB%B0%A5%E...	NaN
4	초독물	서울특별시 동대문구 양장로18길 6 2층	4.2(109) · ₩ · 채식(Vegan)	1	https://www.google.com/maps/place/%EC%B4%B8%EB...	NaN
...
176	저스트텐동 송도트리플스트리트점	인천광역시 연수구 송도동 과학로16번길 33-1 트리플스트리트 A동 별관 2층 20...	4.0(32) · 음식	NaN	https://www.google.com/maps/place/%EC%A0%B0%EC...	1.0
177	저스트텐동 중곡학신도시점	충청북도 음성군 평동면 대하2길 41 105호	4.2(29) · 음식	NaN	https://www.google.com/maps/place/%EC%A0%B0%EC...	1.0
178	저스트텐동 청담점	충청남도 천안시 동남구 청담동 548-1	4.3(10) · 음식	NaN	https://www.google.com/maps/place/%EC%A0%B0%EC...	1.0
179	저스트텐동 포항점	경상북도 포항시 남구 이동 대일로189번길 5-1	4.0(52) · 음식당 및 음식점	NaN	https://www.google.com/maps/place/%EC%A0%B0%EC...	1.0
180	리빙잇 채식뷔페점	경상북도 포항시 남구 이동 대일로189번길 5-1	3.7(3) · 채식(Vegan)	NaN	https://www.google.com/maps/place/%EB%9F%AC%EB...	1.0

식당 데이터프레임

ii. 데이터 전처리

크롤링 하는 과정에서 화면 구성상 별점과 식당의 종류를 함께 크롤링 할 수밖에 없었다. '.'을 기준으로 별점과 종류로 나눠져있어 분리시켜주었다.

별점 데이터를 살펴봤을 때 별점과 별점을 남긴 수가 함께 있어 따로 분리해 주는 작업을 거쳤고, 식당 종류에 대한 데이터를 살펴봤을 때 '채식', '채식(Vegan)' 처럼 같은 종류이지만 다르게 표시된 식당이 있어 통일해 주었다.

```
star_lst=df['star'].str.split('.')
df['review']=star_lst.str.get(0)
df['type']=star_lst.str.get(-1)
df=df.drop('star',axis=1)

for idx in range(df.shape[0]):
    df['type'][idx]=re.sub('^(가-힣)', '',df['type'][idx])
    df['review'][idx]=re.sub('\s','',df['review'][idx])

review_lst=df['review'].str.split('(')
df['star']=review_lst.str.get(0)
df['cnt']=review_lst.str.get(1)
df=df.drop('review',axis=1)
```

주소 데이터에 대해서도 원하는 형태로 맞추기 위해 전처리를 진행하였다. 구분자를 하나로 통일한 후 필요 없는 단어는 지우고, 같은 단어이지만 다르게 처리된 단어들에 대해서도 하나로 통일해 주는 작업을 거쳤다.

```
for idx in range(df.shape[0]):
    df['address'][idx]=re.sub(' ','',df['address'][idx])
    df['address'][idx]=df['address'][idx].replace('KR','')
    df['address'][idx]=re.sub('\s+','',df['address'][idx]).strip()
    df['address'][idx]=df['address'][idx].replace('서울','서울특별시')
    df['address'][idx]=df['address'][idx].replace('특별시특별시','특별시')
```

이후 주소를 자치구까지 나타낸 주소와 세부 주소로 구분하는 작업을 진행해 주었다.

```
def split_address1(address):
    for word in address.split(' '):
        if word in ['서울특별시','경기도','인천광역시','전라남도','충청남도','충청북도']:
            return word

def split_address2(address):
    if '서울특별시' in address or '인천광역시' in address:
        for word in address.split(' '):
            if '구' in word:
                return word
    elif '경상북도' in address:
        for word in address.split(' '):
            if '구' in word:
                return word
            elif '시' in word:
                return word
    elif '경기도' in address or '전라남도' in address or '충청남도' in address or '충청북도' in address:
        for word in address.split(' '):
            if '시' in word or '군' in word:
                return word
    else:
        return address
```

```
def split_address3(address):
    sub=[]
    if '서울특별시' in address or '인천광역시' in address:
        for word in address.split(' '):
            if '구' in word:
                word=''
                if len(word)>0:
                    sub.append(word)
    elif '경상북도' in address:
        for word in address.split(' '):
            if '구' in word:
                word=''
            elif '시' in word:
                word=''
                if len(word)>0:
                    sub.append(word)
    elif '경기도' in address or '전라남도' in address or '충청남도' in address or '충청북도' in address:
        for word in address.split(' '):
            if '시' in word or '군' in word:
                word=''
                if len(word)>0:
                    sub.append(word)
```

```

result=' '
for word in sub:
    if word in ['서울특별시', '경기도', '경상북도', '전라남도', '충청남도', '충청북도']:
        pass
    else:
        result=result+' '+word
return result.strip()

```

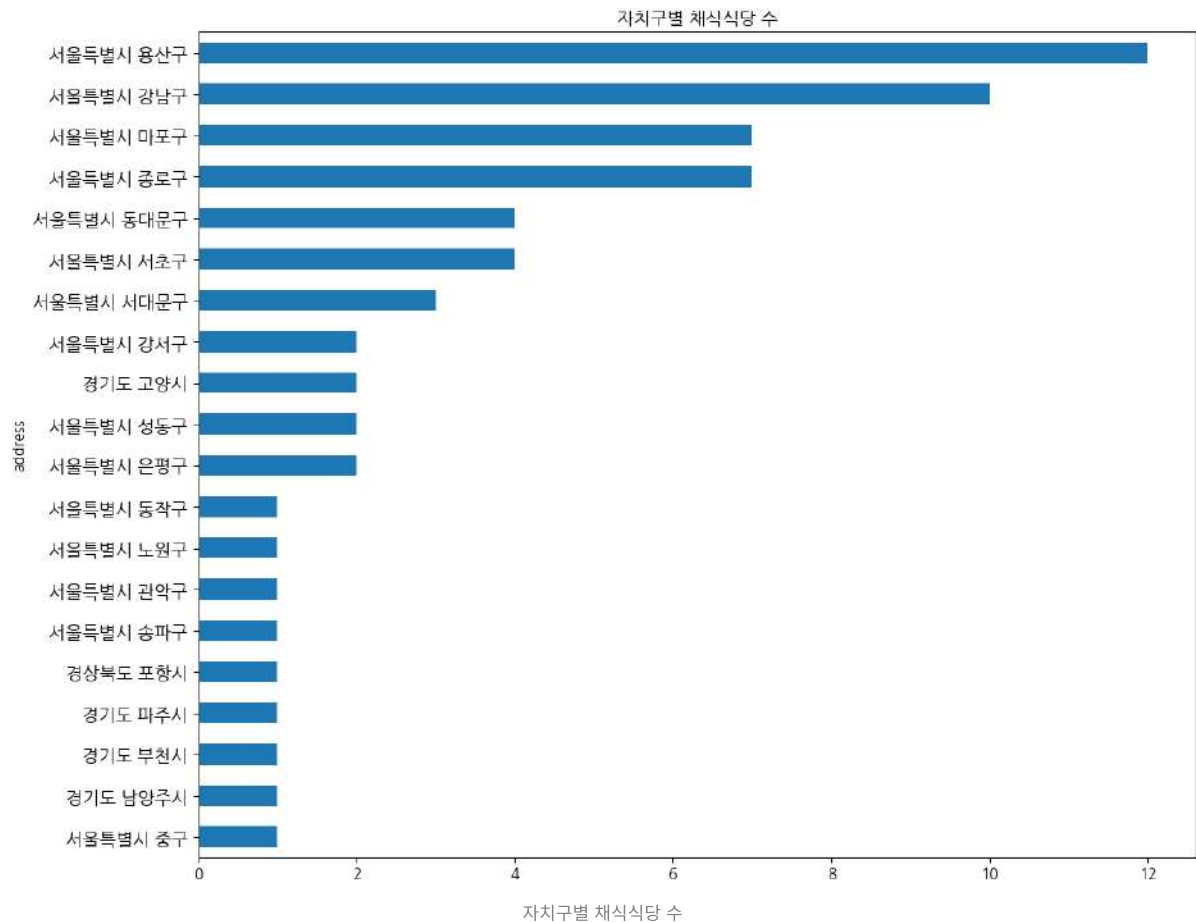
전처리를 모두 거친 식당 데이터의 최종 데이터 프레임이다.

	name	vegan	url	pesco	type	star	cnt	address1	address2	sub_address	address	latitude	longitude
0	라브 더 비건즈	1	https://www.google.com/maps/place/%EB%9F%AC%EB...		채식	리뷰 없음	0	경기도	남양주시	호평동 642-2	경기도 남양주시	37.635940	127.216505
1	담향비빔국수	1	https://www.google.com/maps/search/%EB%A7%9D%E...		음식점	4.2	20	경기도	남양주시	호평동 늘운동로14번길 29	경기도 남양주시	37.635940	127.216505
2	마히나 비건 테이블 Mahina Vegan table	1	https://www.google.com/maps/place/%EB%A7%88%ED...		채식	4.2	43	서울특별시	강남구	논현로175길 75 2층	서울특별시 강남구	37.517700	127.047300
3	밥풀꽃	1	https://www.google.com/maps/search/%E9%80%A5%E...		채식	리뷰 없음	0	서울특별시	노원구	상계동 966-1번지 1층 중현빌딩	서울특별시 노원구	37.654000	127.056700
4	초록돌	1	https://www.google.com/maps/place/%EC%84%88%EB...		채식	4.2	109	서울특별시	동대문구	망우로18길 6 2층	서울특별시 동대문구	37.574198	127.039509
...
176	저스트텐동 송도트리 물스트리트점		https://www.google.com/maps/place/%EC%A0%80%EC...	1.0	음식	4.0	32	인천광역시	연수구	인천광역시 송도동 과학로16번길 33-1 드림스트리트 A동 별관 2층 207-3호	인천광역시 연수구	37.409800	126.678700
177	저스트텐동 중북혁신 도시점		https://www.google.com/maps/place/%EC%A0%80%EC...	1.0	음식	4.2	29	충청북도	음성군	영동면 태하2길 41 105호	충청북도 음성군	36.940000	127.690600
178	저스트 텐동 청담점		https://www.google.com/maps/place/%EC%A0%80%EC...	1.0	음식	4.3	10	충청남도	천안시	동남구 청담동 548-1	충청남도 천안시	36.815028	127.114065
179	저스트텐동 포항점		https://www.google.com/maps/place/%EC%A0%80%EC...	1.0	일식당및 일명식집	4.0	52	경상북도	포항시	이동 대이로189번길 5-1	경상북도 포항시	36.018932	129.342938
180	리빙컷 채식뷔페점		https://www.google.com/maps/place/%E6%9F%AC%EB...	1.0	채식	3.7	3	경상북도	포항시	이동 대이로189번길 5-1	경상북도 포항시	36.018932	129.342938

181 rows × 13 columns

식당 데이터 전처리 완료 데이터프레임

이렇게 수집한 채식 및 페스코 식단이 가능한 식당의 업종에는 어떤 종류가 있는지 그래프를 그려보았다.



사람들이 많이 모이는 지역이 서울특별시 용산구와 강남구에 많은 식당들이 위치해 있었고, 그 외 경기도 지역에서도 종종 위치해 있는 것을 확인할 수 있다.

아무래도 구글에서 크롤링 하는 과정에서 현재 위치에 기반한 식당들을 위주로 검색이 진행되다 보니 서울과 경기도에 치중된 결과를 가져온 경향도 있어 보인다. 다양한 지역을 중심으로 검색을 진행한다면 더 많은 식당과 더 다양한 위치의 식당 데이터도 수집할 수 있을 것이라고 기대된다.

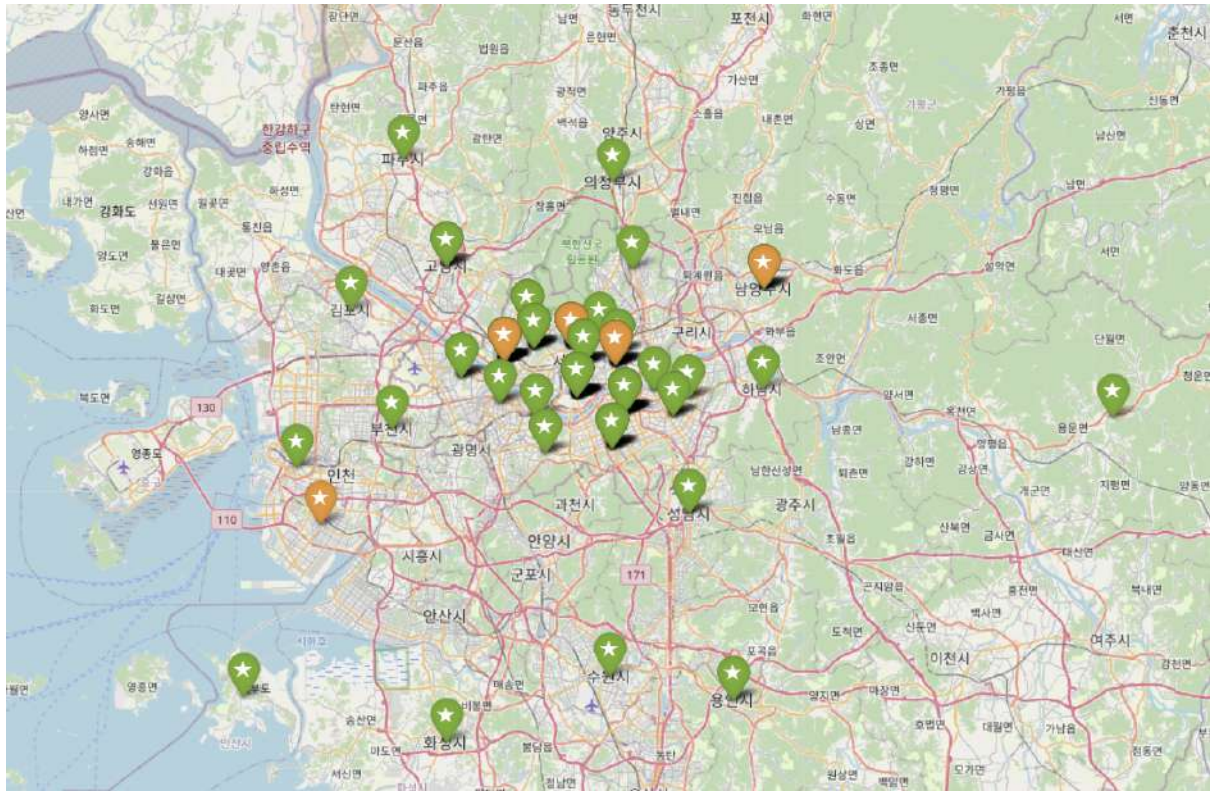
iii. 지도 시각화

folium 라이브러리를 사용해 식당데이터를 다양한 방법으로 시각화 해보았다.

지도에 식당의 위치를 마크로 표시하기 위해 주소를 사용해 위도와 경도를 구해 저장해 두었다.

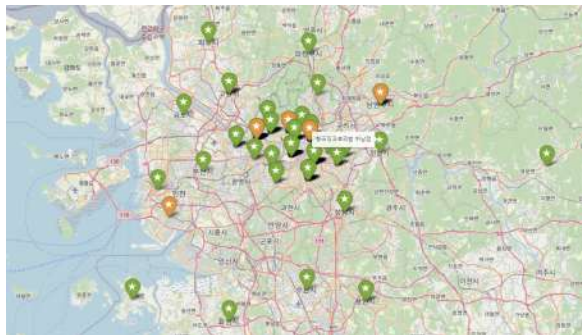
```
for i in df['address'].tolist():
    latitude.append(geocoding(i)[0])
    longitude.append(geocoding(i)[1])
```

우선 식당의 위도와 경도를 활용해 지도 위에 식당의 위치를 마크로 표시하였다. 비건 식당이 가능한 식당은 초록색으로, 페스코 식당이 가능한 식당은 주황색으로 표시하였다.

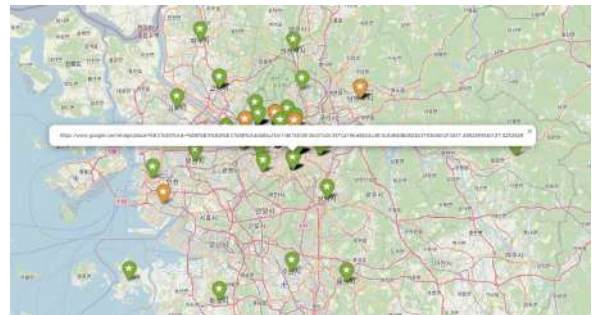


식당 데이터 시각화1-1

마크로 표시된 곳 위에 마우스를 올리면 식당의 이름이 나타나고, 마크를 클릭하면 식당에 대한 자세한 정보와 리뷰를 볼 수 있는 url이 뜬다.



식당데이터 시각화1-2

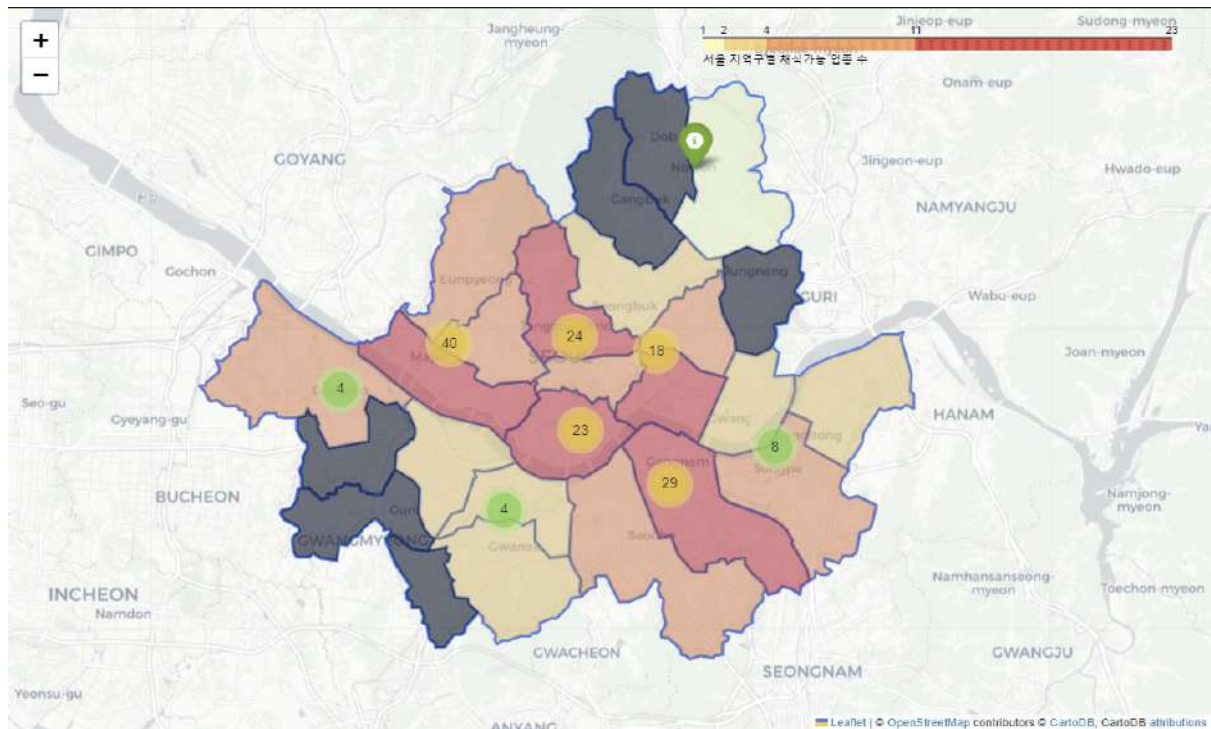


식당데이터 시각화1-3

html 파일을 활용한다면 확대와 축소 및 이동도 가능하다.

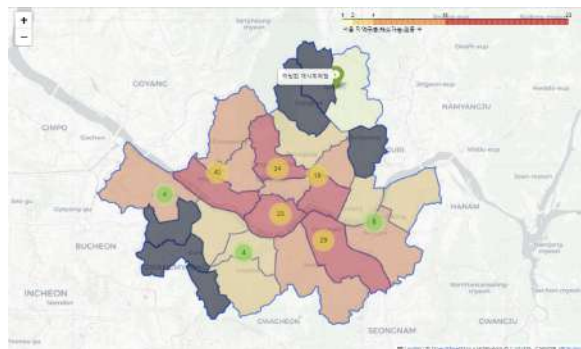
<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/db784c5a-af65-4ccc-855c-2ef57833fcb0/map1.html>

식당 데이터 중 서울에 위치한 식당에 대해서만 시각화를 진행해 보았다. 어느 자치구에 비건 가능 식당이 몇 개나 존재하는지를 한눈에 보기 쉽게 정리하였다.

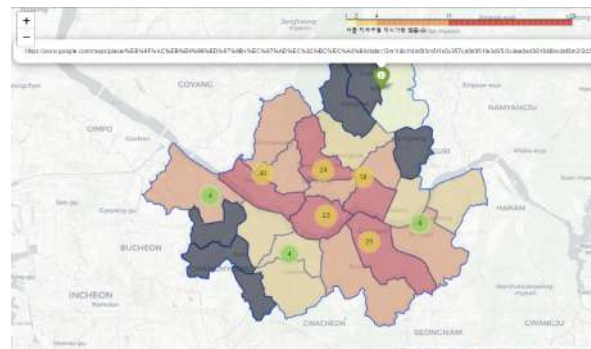


식당데이터 시각화2-1

마크로 표시된 곳 위에 마우스를 올리면 식당 이름이, 마크를 클릭하면 식당에 대한 자세한 정보와 리뷰를 볼 수 있는 url을 보여준다.



식당 데이터 시각화2-2



식당 데이터 시각화 2-3

마찬가지로 html을 이용하면 확대 및 축소와 이동이 가능하다.

<https://s3-us-west-2.amazonaws.com/secure.notion-static.com/99801214-0a0c-48ad-b340-79264c0af80a/map3.html>

6. 서비스 기획

a. 비건 서비스 사전 조사

- 채식한끼 피드/레시피/콘텐츠/식당검색/스토어/챌린지
- 비거너 챌린지
- 비건로드 비건식당검색/블로그리뷰
- 비니티 커뮤니티/제품검색/비건제품확인/온라인몰

- **채크인** 챌린지/동네기반모임/식당카페
- **비건북** 비건음식쇼핑몰
- **브릿지** 채식단계선택/채식피드공유
- **비건어게인** 비건제품확인
- **Ralphy** 외국어플/크루얼리티확인/국내제품확인불가

b. 비건 어플리케이션 기획

이름

- 베지메이트 Veggie Mate

Veggie
Mate

세계관

- 미션 : 사람들의 비건 생활을 돕는다
- 비전 : 모두의 일상에 비건을!

가치 제안 캔버스

1. 고객 프로필

비건을 시작 하려는 대학생	25세. 환경문제에 대한 의식이 높아져 비건을 시작해보려함.
고객 과업	- 비건 식단을 시도한다 - 비건에 대한 정보를 얻는다
고충	- 어떤 것이 비건 제품인지 확인하기 어렵다 - 비건을 어떻게 시작해야할지 모른다
이득	- 환경문제에 대한 적극적인 참여
락토 베지테리언 생활을 유지 중인 직장인	32세. 건강과 동물권에 대한 관심으로 락토 베지테리언 생활을 유지 중임.
고객 과업	- 락토 베지테리언 생활 유지 혹은 더 높은 단계의 베지테리언에 도전 - 제품의 알러지 성분의 함유 여부 확인
고충	- 기존 비건 어플로는 락토에 해당하는 제품인지 확인이 어렵다 - 기존 비건 어플에서는 일일이 알러지 성분을 확인해야 한다
이득	- 건강한 생활과 신체

2. 가치 맵

비건을 시작 하려는 대학생	25세. 환경문제에 대한 의식이 높아져 비건을 시작해보려함.
프로덕트 및 서비스	- 비건 식단에 대한 정보 제공(레시피, 식당, 식품 등) - 비건에 대한 기본적인 정보를 알려주는 기능 제공
고충 해소책	- 성분 분석을 통해 비건 제품을 알려주는 서비스 - 비건 생활 챌린지를 통한 입문 서비스

비건을 시작 하려는 대학생	25세. 환경문제에 대한 의식이 높아져 비건을 시작해보려함.
가치 창출책	- 본인의 비건 생활로 인해 환경 문제에 어떻게 기여했는지를 알려주는 기능
락토 베지테리언 생활을 유지 중인 직장인	32세. 건강과 동물권에 대한 관심으로 락토 베지테리언 생활을 유지 중임.
프로덕트 및 서비스	- 다양한 비건 단계에 도전할 수 있는 챌린지 서비스 - 설정한 알러지 성분의 함유 여부를 알려주는 서비스
고충 해소책	- 제품이 비건의 어느 단계에 해당하는 제품인지 알려준다 - 설정한 알러지 성분의 함유 여부를 알려주는 서비스
가치 창출책	- 비건 생활을 기록하고 건강의 변화를 함께 보여주는 서비스 제공

린 캔버스

항목	설명
과제	- 시중의 비건 어플의 성분 분석 시스템은 완전 비건인을 대상으로 서비스를 제공하여 오보, 락토오보, 페스코 등 부분적인 채식을 하는 경우 확인이 번거롭다. - 시중의 비건 어플은 직접 제품 이름을 검색하여 제품의 비건 여부를 확인해야하므로 제품이 등록되어 있지 않은 경우 확인이 어렵다. - 시중의 비건 어플은 완전히 채식을 하는 경우를 상정하고 서비스를 제공하여 비건을 가볍게 시작하기에 장벽이 높다. - 시중의 비건 어플들은 기능이 분산되어 있어 여러 어플을 사용해야하는 번거로움이 있다. - 시중의 비건 어플들은 대부분 '채식'에 방점을 두고 있어 비건 '생활'의 영역까지는 커버를 하지 못한다. 기존의 대안 → 비건 어플리케이션 - 비니티 - 채식한끼 - 비거너 - 비건로드 - 채크인 - 비건룩 - 브릿지 - 비건어게인 - Ralphy → SNS → 검색 엔진(구글 등)
솔루션	- 제품의 성분표 사진을 찍으면 이를 직접 분석하여 비건 단계별로 분류해줄 수 있는 기능. - 다양한 비건 단계별을 아우르는 정보(레시피, 식당 정보 등) 제공. - 비건 생활에 도움이 되는 기능을 어플 하나로 통합한다. - 비건 생활 챌린지 기능을 통해 다양한 비건 생활을 경험해보고 지속할 수 있는 동기를 부여한다. - 식품 뿐만 아니라 화장품의 동물성 성분 여부도 확인할 수 있도록 하여 '생활'의 영역까지 확장한다.
측정 가능한 주요 지표	- 신규 유저 수 - 유저 이탈률
프로덕트가 고객에게 주는 독특한 가치	- 비건을 시작하는 문턱을 낮추고 지속할 수 있는 동기를 부여해준다. 높은 수준의 콘셉트 → 모두의 일상에 비건을!
경쟁 우위	- 제품에 관계 없이 성분을 분석할 수 있음 - 식품과 화장품 모두 아우르고 있음
유통채널	- 앱스토어 - 광고
고객군	- 비건을 막 시작하려는 스타터들 - 본인이 지향하는 비건 단계를 명확히 알고, 유지해나가는 사람들 오픈 어답터 → 부분적으로 채식을 하는 유저들
비용 구조	- 초기 투자 비용 : 개발 비용 - 고정비 : 운영 비용 - 변동비 : 마케팅 비용
수입원	- 비건 제품을 생산하는 기업과의 제휴 광고 - 어플 광고 수입 - 유저 대상 멤버십 : 광고 제거 및 레시피 선공개, 체험 기회 제공 등

과제 - 식품의 비건 여부를 성분 분석 시스템은 완전 비건으로 대상으로 서비스를 제공하여 정보, 특효효과, 특효효과 등 부분적인 제식을 하는 경우 확인이 필요하다. - 식품의 비건 여부를 직접 제품 이름을 검색하여 식품의 비건 여부를 확인해야하므로 제품의 정확도가 있지 않은 경우 확인이 어렵다. - 식품의 비건 여부를 완전한 제식을 하는 경우를 운영하고 서비스를 제공하여 비건들 가정에 시작하기가 어려워진다. - 식품의 비건 여부를 직접 제품 이름을 검색하여 식품의 비건 여부를 확인해야하므로 제품의 정확도가 있지 않은 경우 확인이 어렵다. - 식품의 비건 여부를 직접 제품 이름을 검색하여 식품의 비건 여부를 확인해야하므로 제품의 정확도가 있지 않은 경우 확인이 어렵다.	솔루션 - 제품적 성분표 사진을 찍으면 이를 직접 분석하여 비건 단계별로 분류해 줄 수 있는 기능. - 다양한 비건 단계별을 아우르는 정보(비건, 비건 정도 등) 제공. - 비건 생활에 도움이 되는 기능을 제품 하나로 통합한다. - 비건 생활 정보와 기능을 통해 다양한 비건 생활을 경험해보고 지속할 수 있는 동기를 부여한다. - 식물 뿐만 아니라 화장품의 동물성 성분 여부도 확인할 수 있도록 하여 생활의 편리까지 제공한다.	프로덕트가 고객에게 주는 독특한 가치 - 비건을 시작하는 문턱을 낮추고 지속할 수 있는 동기를 부여해준다.	경쟁 우위 - 제품에 관계 없이 성분을 분석할 수 있음 - 식품과 화장품 모두 아우르고 있음	고객군 - 비건을 막 시작하려는 스타터들 - 본인이 지향하는 비건 단계를 명확히 알고, 유지해나가는 사람들
기존의 대안 비건 어플리케이션 - 비니더 - 제식함 - 비거니 - 비건로드 - 제크인 - 비건북 - 브릿지 - 비건어제전 - Ralphy SNS 검색 엔진(구글 등)	측정 가능한 주요 지표 - 신규 유저 수 - 유저 이탈률	높은 수준의 콘셉트 모두의 일상에 비건을!	유통채널 - 앱스토어 - 광고	얼리 어답터 부분적으로 제식을 하는 유저들
비용 구조 - 초기 투자 비용 : 개발 비용 - 고정비 : 운영 비용 - 변동비 : 마케팅 비용		수입원 - 비건 제품을 생산하는 기업과의 제휴 광고 - 어플 광고 수입 - 유저 대상 멤버십 : 광고 제거 및 레시피 선공개, 체험 기회 제공 등		

서비스 기획서

1. 서비스 개요

- 서비스명: 베지메이트
- 서비스 목적: 비건 생활을 시작하고 지속하는 것을 도와주는 정보 및 서비스 제공
- 대상 사용자: 비건 생활을 시도해보고 싶은 사용자, 비건 생활을 재미있게 지속하고 싶은 사용자

2. 주요 기능

a. 성분 분석 및 추천 기능

- 식품 성분 분석 및 추천 기능: 사용자가 검색한 식품의 성분을 분석하여 어느 단계의 비건까지 섭취 가능한지 분류해 알려주는 기능, 분석한 성분을 바탕으로 유저가 원하는 비건 단계의 제품을 추천해주는 기능
- 화장품 성분 분석 및 추천 기능: 사용자가 검색한 화장품의 성분을 분석하여 동물성 성분의 포함 여부를 확인 하고 알려주는 기능, 분석한 성분을 바탕으로 유저가 원하는 카테고리, 비건 단계의 제품을 추천해주는 기능
- 알러지 성분 판별 기능: 사용자가 가진 알러지(사전 입력)가 식품에 포함되어있는지 알려주고, 화장품의 경우 알러지 유발 성분이 무엇이 있는지 알려주는 기능



b. 리뷰 제공 및 분석 기능

- 식품, 화장품의 리뷰를 제공해주고 분석해 주요 키워드를 알려주는 기능



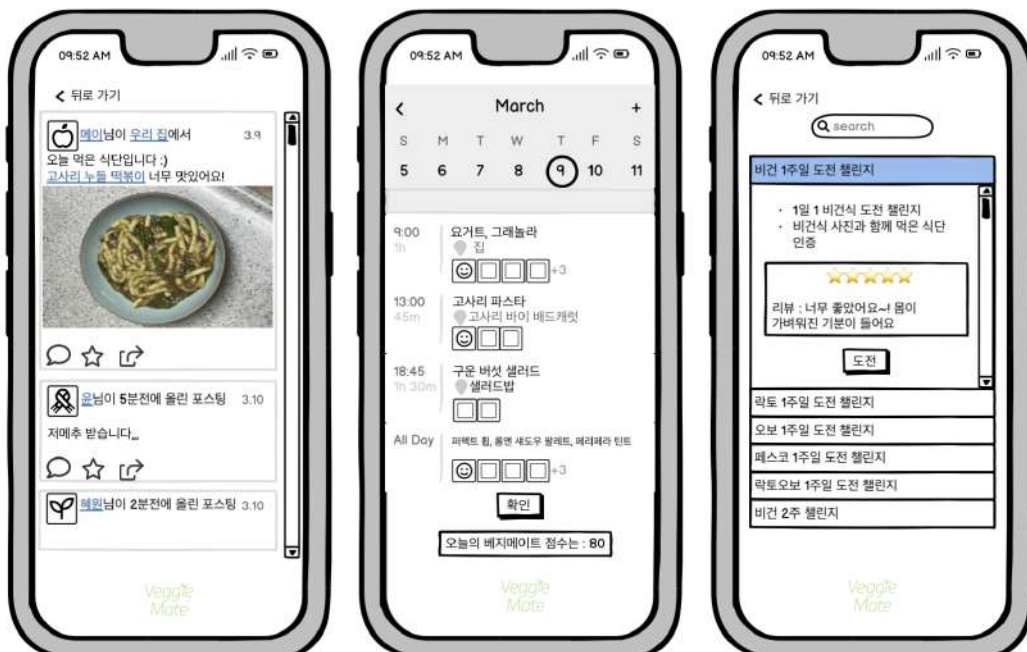
c. 비건 식당 정보 제공 기능

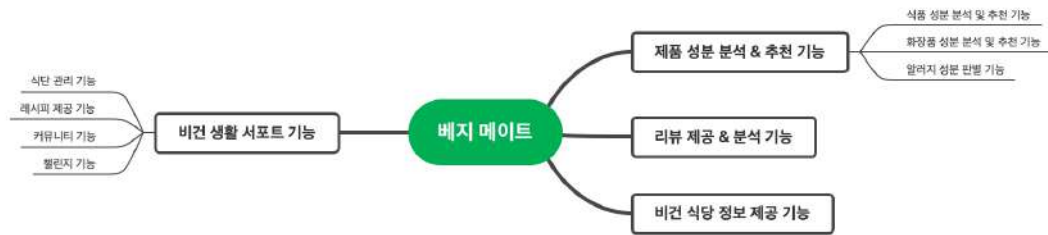
- 주소를 입력하고 비건 단계를 설정하면 근처의 섭취 가능 식당을 알려주고 지도에 시각화하여 보여주는 기능



d. 비건 생활 서포트 기능

- 식단 관리 기능: 사용자가 섭취한 음식을 기록하여 영양 성분을 확인하고, 식단을 추천해주는 기능
- 레시피 제공 기능: 다양한 단계의 비건 요리 레시피를 제공하는 기능
- 커뮤니티 기능: 비건 생활에 대한 정보 및 경험을 공유하는 커뮤니티 기능 제공
- 챌린지 기능: 다양한 비건 단계에 대한 일주일~한달 챌린지를 제공하여 한끼~세끼를 각 비건 단계별로 시도해보고 인증하며 비건 생활에 익숙해지도록 도와주는 기능





3. 필요 기능

- 카메라 기능: 사용자가 촬영한 성분표를 바탕으로 비건 단계를 분석함.
- 휴대폰 내 건강기록 기능 : 휴대폰에 등록된 사용자의 건강 상태와 연동하여 유저의 식단 분석 및 건강 상태 분석.

4. 수익 모델

- 광고 수익: 비건 제품 관련 광고를 게재하여 수익 창출.
- 어플 내 광고 수익 : 어플 사용시 광고가 노출되도록 하여 광고 수익 창출.
- 유료 기능: 멤버십 서비스(월결제, 연결제)를 제공하여 광고 제거 기능, 제휴 비건 제품 체험 서비스, 비건 레시피 선공개 기능 등을 제공함.

7. 결론

a. 차별점

1. 비건 단계별 성분 분석 서비스를 제공하여 다양한 비건인들의 수요를 충족한다.
2. 사진으로 성분을 바로 분석할 수 있어 사용자가 쉽게 비건 제품을 파악할 수 있도록 도왔다.
3. 성분 분석 기능을 응용해 알러지 및 글루텐 성분을 알려주는 기능을 추가해 유저의 불편을 최소화 한다.

b. 한계점 및 발전 방향

1. 화장품의 동물 실험 여부는 일일이 회사에 전화해서 물어보아야해서 이번 프로젝트에서는 동물성 성분만 분석, 분류하는 방식으로 진행하였다. 실제 서비스를 런칭할 경우 동물성 실험 여부도 확인할 필요가 있다.
2. 식품 성분이 비건 단계별로 분류된 자료가 없어 성분 리스트를 직접 만들고 분류하여 전처리를 진행하였는데, 분류 과정에서 개개인의 주관이 개입되었을 가능성이 있어 전문성이 떨어진다. 실제로 서비스를 런칭할 경우 식품 관련 전문가의 자문이 필요할 것으로 보인다.
3. 이미지에서 텍스트를 추출할 때에 정확도가 낮으므로 추후 서비스를 런칭할 경우 텍스트 인식 정확도를 높일 필요가 있다.
4. 시간이 부족하여 리뷰 분석이 미흡하였다. 추후 런칭할 경우 상품에 대한 키워드를 추출하고 분석하여 필터링하는 기능이 있다면 유저가 사용하기에 더 편리할 것으로 보인다.
5. 시간이 부족해 서비스 기획에만 그친 부분이 많다. 구체적인 구현을 해보았다면 더 좋았을 것 같다.
6. 마케팅적으로 어떻게 홍보하거나 대중화 시킬지 고민해보지 못한 점이 아쉽다.

8. 개발 후기 및 느낀 점

	개발 후기 및 느낀점
박정호	프로젝트를 진행하면서 어느 부분이 미흡한지 알게 되었고, 부족한 점을 어떻게 채워야 할 지 생각하게 되었다. 팀원분들이 다 잘해주셔서 프로젝트를 잘 마무리 할 수 있었고, 앞으로 또 다른 프로젝트를 진행한다면 어떻게 해야 할 지 배워가는 시간이 된 것 같다.
변재윤	데이터도 프로젝트 자체가 처음이라 시간도 오래 걸리고 어떤 방향성을 가지고 진행되는지조차 몰랐는데 좋은 팀원들을 만나 끝까지 진행 될 수 있음이 행운이라 생각합니다. 아는것보다 모르는게 더 많은 상황에서 아주 오래 헤매면서 코드를 조금이나마 작성할 수 있게 다른 작업을 다 도맡아서 해주신 팀원분들께 참 죄송하고 프로젝트 전보다 발전했다고 자신할 수 없지만 많은 걸 배웠음은 확실한거 같습니다.
성지영	1달 이상의 긴 프로젝트가 처음이라 어떻게 이끌어갈 지, 일정 관리는 어떻게 하고 팀원들 간의 의사 소통을 조율하는 것은 어떻게 해야하는 지 몰라서 어려움이 많았다. 팀원들의 배려와 적극적인 참여로 프로젝트를 무사히 마무리할 수 있어 정말 기쁘고, 이번 기회를 통해 데이터 분석 및 개발 뿐만 아니라 프로젝트의 일정과 관리에 대해서도 많이 배웠다.
이혜원	크롤링, 텍스트 추출등을 통해 얻은 비정형화된 텍스트를 활용해서 분석을 진행하다보니 어려움도 많았지만 다양한 전처리 과정을 거쳐 원하는 결과물을 이끌어낼 수 있는 좋은 경험의 시간이었다. 또한 결과물을 사용하여 다양한 시각화 기법을 적용해 볼 수 있어서 좋았다.

9. 레퍼런스

1. 식품안전나라 <https://foodsafetykorea.go.kr/main.do>
2. 쿠팡 <https://coos.kr/products?page=1>
3. 한국 비건 인증원 <http://vegan-korea.com/production>
4. 대한 화장품 협회 <https://kcia.or.kr/cid/main/>