데이터 분석 및 활용
캡스톤 디자인 과제

-반려묘 사료 영양 성분 분석-

1716853 김주은
1614072 박영미

목차

.

<역할분담>

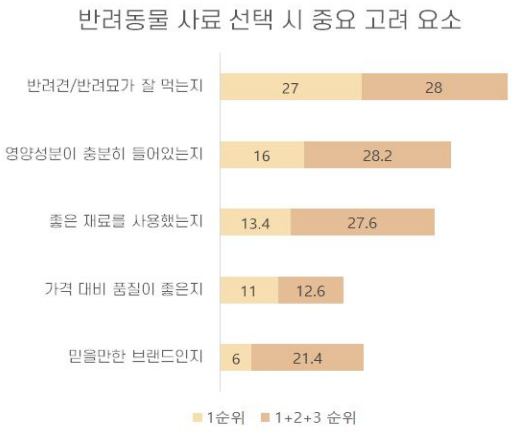
김주은: 피피티, 보고서, 영양 성분 함량별 분석(클러스터링, 워드 카운트, 상관관계분석)

박영미: 보고서, 크롤링생애단계별 분석(유사도 계산, 워드카운트), 리뷰분석, 원료점수-리뷰평점 상관관계분석

1. 서론

1.1 분석 주제 선정 배경

최근 '펫팸족'의 증가로 펫코노미가 블루오션으로 떠올랐다. 펫팸족이란 '애완동물(pet)'과 '가족(family)'의 합성어로 반려동물을 가족처럼 생각하는 사람들을 의미한다. 2020년 반려동물 시장 규모는 5조8000억원으로 전망된다. 특히 글로벌 시장 조사 회사 '유로모니터'에 따르면, 2020년 기준 국내 펫푸드 시장 규모(소비자가 기준, 개&고양이)는 약 1조 2650억원에 이른다.



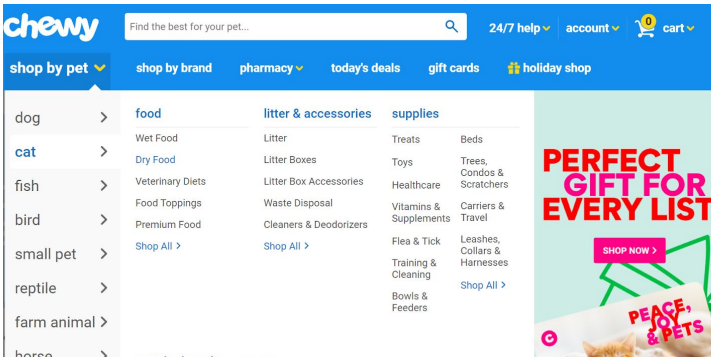
반려동물 사료 선택 시 여러 요소를 고려하는데, 왼쪽 표를 보면 반려견/반려묘가 잘 먹는지, 영양성분이 충분히 들어있는지, 좋은 재료를 사용했는지, 가격 대비 품질이 좋은지, 믿을만한 브랜드인지를 고려한다는 것을 알 수 있다. 하지만 현재 제공되고 있는 사료 추천 시스템은 단순히 사용자들의 구매정보에 기반한 추천이 이루어지는 경우가 많다. 또 반려동물에게 적합한 원료인지 판단하기가 어려워 고양이 사료 등급 정도로 파악하지만, 이것마저도 충분하지 않다고 생각하는 소비자들이 81%였다. 하지만 반려묘의 품종, 연령, 크기, 신체 상태, 질병 등에 따라 섭취해야 할 영양성분이 달라지기 때문에 필요한 영양성분을 포함한 사료인지 판단하는 것이 중요하다.

1.2 분석 목적

본 프로젝트에서는 반려동물 중에서 반려묘의 사료를 대상으로 영양 성분 함량, 원료 구성, 연령을 고려하여 분석하고 원료 구성에 기반한 추천 시스템을 제안한다. 시중에서 유통되고 있는 578개의 고양이 건식사료의 데이터를 수집하고, 군집분석과 성분별 상관관계를 분석하여 유사성을 측정한다. 본 프로젝트의 분석 결과는 소비자들이 사료 선택 시 사료 비교에 도움을 주고 사료 추천시스템 같은 맞춤형 서비스에 활용할 수 있다.

1.3 분석 데이터

1) 'Chewy' 사이트에서 건식 사료 정보만 추출

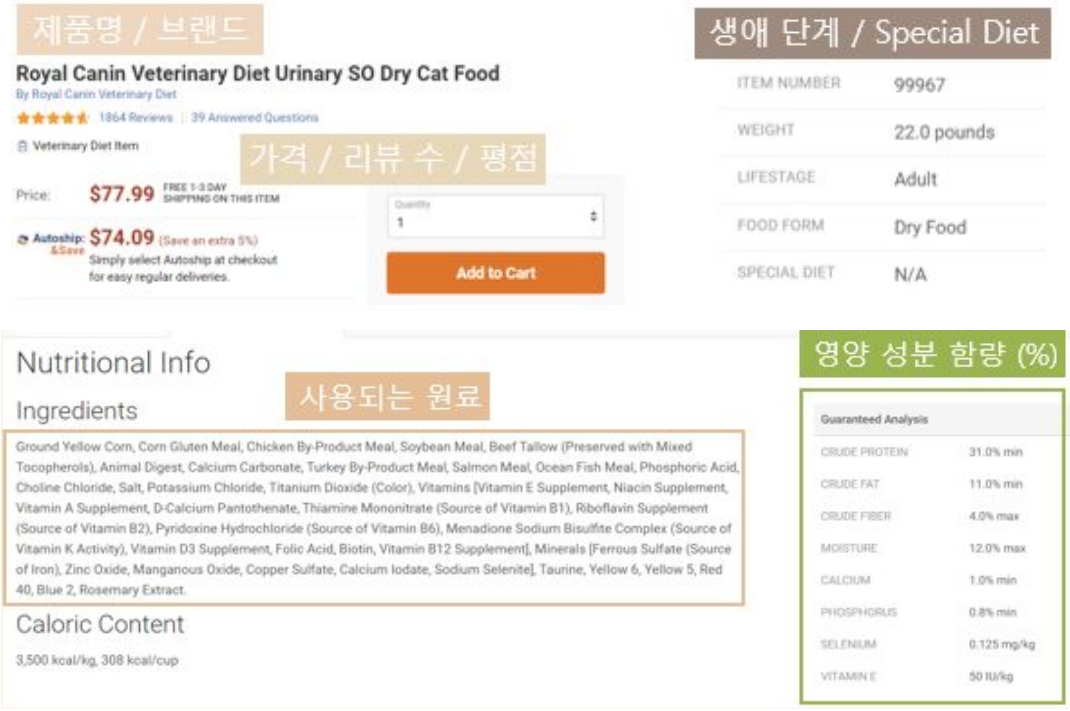


사이트: https://www.chewy.com/ 카테고리: Cat › Food › Dry Food

2) 데이터 수집 (사료 582개)

① 제품명 ②브랜드 ③가격 ④연령대 ⑤Special diet(기능)

⑥리뷰 ⑦리뷰 평점 ⑧원료 영양성분 함량 (조단백, 조지방, 조섬유)



3) 데이터 크롤링 및 전처리

```
product = soup.select('#product-title > h1')[0].text  # 제품명
brand = soup.select('#product-subtitle > a > span')[0].text  # 브랜드명
price = soup.select('#pricing > ul > li.our-price > p.price > span.ga-eec__price')[0].text  # 가격
try:
    life_stage = soup.find('div', text=re.compile('Lifestage')).findNext('span').text  # 생애단계
except:
    life_stage = ''
```

```
data = pd.DataFrame(table) #dataframe형태로 변경
data.head() #상위정보 출력
data.columns = ['product', 'brand', 'price', 'life_stage', 'special_diet', 'review_num', 'review_star',
                'ingredients', 'protein', 'fat', 'fiber', 'moisture']  # 테이블 열 이름 변경
```

beautiful soup, selenium을 이용해 chewy 에서 제공하고 있는 사료 데이터를 웹 크롤링 했다.

582개 사료 정보를 수집하여 csv 파일 형태로 저장했다.

4) 전처리 과정

a. 양성분 4개 중 2개 이상 없는 항목 제거

b. ‘life_Stage’열 - 비어있는 항목 처리

c. ‘special_diet’ 열 - N/A → Normal 로 변경 (특별한 기능이 없는 사료)

d. ‘fiber; 열 - 결측치 2개 발견 → 항목 제거, 분석에서 제외

2. 분석 시나리오 및 결과

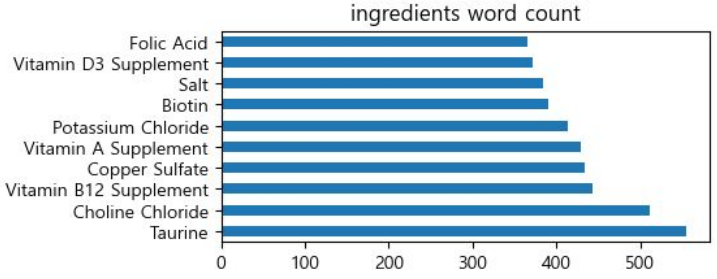
2.1 연령대(Kitten, Adult, Senior)별 분석

1) 전체 원료 개수 Wordcount

```
catfood_data = pd.read_csv('adult_data.csv', encoding='CP949')
data1 = catfood_data['ingredients']
# print(all[0].split(','))
items = []  # 모든 원료
for i in data1:
    item = i.split(',')  # ,로 여러 원료 구분x
    for j in item:
        items.append(j.strip())  # 전체리스트에 추가

print('중복 고려안한 모든 원료 개수: ', len(items)) #중복을 고려한 원료 개수

items_lower = []
for k in items:
    items_lower.append(k.lower().strip())
items_lower = pd.DataFrame((items_lower))
print('중복을 고려한 원료 개수: ', len((items_lower))) # 중복을 고려한 원료 개수
```




```
원료 개수 카운트
Taurine                      555
Choline Chloride             512
Vitamin B12 Supplement       443
Copper Sulfate               434
Vitamin A Supplement         429
Potassium Chloride           413
Biotin                       390
Salt                         384
Vitamin D3 Supplement        372
Folic Acid                   366
Calcium Iodate               354
Thiamine Mononitrate         351
Sodium Selenite              336
Riboflavin Supplement        332
Pyridoxine Hydrochloride     329
Ferrous Sulfate              319
Calcium Carbonate            298
Natural Flavor               296
Manganese Sulfate            282
Vitamin E Supplement         281
```

→ 닭고기, 참치, 연어 등 사료에 많이 쓰이는 원료들이 주를 이룰 것으로 예상했으나, Taurine, Choline Chloride, Vitamin 등의 원료가 상위를 차지했다.

2) 연령대(Kitten - Adult) / (Adult - Senior) / (Senior - Kitten) 별로 원료 구성에서 차이가 존재하는지 파악하기 위해 3가지 척도로 유사도 계산

```
#euclidean distance
def Dist(x,y):
    return np.sqrt(np.sum((x-y)**2))

#cos similarity
def CosSim(a, b):
    return np.dot(a,b)/(np.sqrt(sum(np.square(a))) * np.sqrt(sum(np.square(b))))

#jaccard_similarity
def JaccardSim(a,b):
    doc1 = set(a)
    doc2 = set(b)
    return len(doc1&doc2) / len(doc1|doc2)
```

	유클리드 거리	코사인 유사도	자카드 유사도
Kitten – Adult	0.7745	0.8393	0.3522
Adult - Senior	1.3924	0.7782	0.2400
Senior – Kitten	1.4983	0.6602	0.5022

→ 코사인 유사도가 1보다 작은 것을 통해 그룹마다 원료 구성이 유사하지 않고,
자카드 유사도 결과를 통해 원료 구성에 차이가 있음을 파악했다.

→ Word count를 통해 각 연령대 별로 특이적으로 등장하는 원료가 있는지 분석했다.

3) 각 연령대 그룹에서 특이적으로 많이 나타나는 원료를 파악하기 위해 tf-idf 실시

Adult 사료 수가 507개로 Adult에 값이 치중될 것을 고려해 Boolean 형태로 data 변환 후 Idf 값과 곱한 결과, 총 원료 개수 1424개, 공통으로 가지고 있는 원료는 322개, kitten에만 있는 원료는 27개, adult에만 있는 원료는 791개, senior에만 있는 원료는 35개로 확인되었다.

```
                                                  kitten  adult  senior
Taurine                                                1      1       1
Choline Chloride                                       1      1       1
Vitamin B12 Supplement                                 1      1       1
Vitamin A Supplement                                   1      1       1
Copper Sulfate                                         1      1       1
...                                                  ...    ...     ...
Chicken Meal (Source of Glucosamine and Chondro...     0      0       1
Soybean Oil (Preserved With Mixed Tocopherols)         0      1       0
L-Ascorbyl- 2-Polyphosphate (a Source of Vitami...     0      1       0
Dried Lactobacillus Casei                              0      1       0
Fermentation Solubles                                  0      1       0
```

```
#idf
idf_list = []
for k in range(len(kitten_items)):
    idf_data = itotal[:,k] != 0 # 단어를 가지
    idf_data = idf_data.astype(np.int).sum()
    idf_data = math.log10(3/idf_data)
    idf_list.append(idf_data)  # 전체리스트에

idf = pd.DataFrame(idf_list)
bool_total['idf'] = idf_list
print((idf == 0).astype(np.int).sum())
print(bool_total)


#tf idf
data = []
for i in range(3):
    t =(((bool_total.iloc[:,i].mul(bool_total
    t = t[t != False].index.values
    data.append(t)
    print(t)
    print(len(t))
```

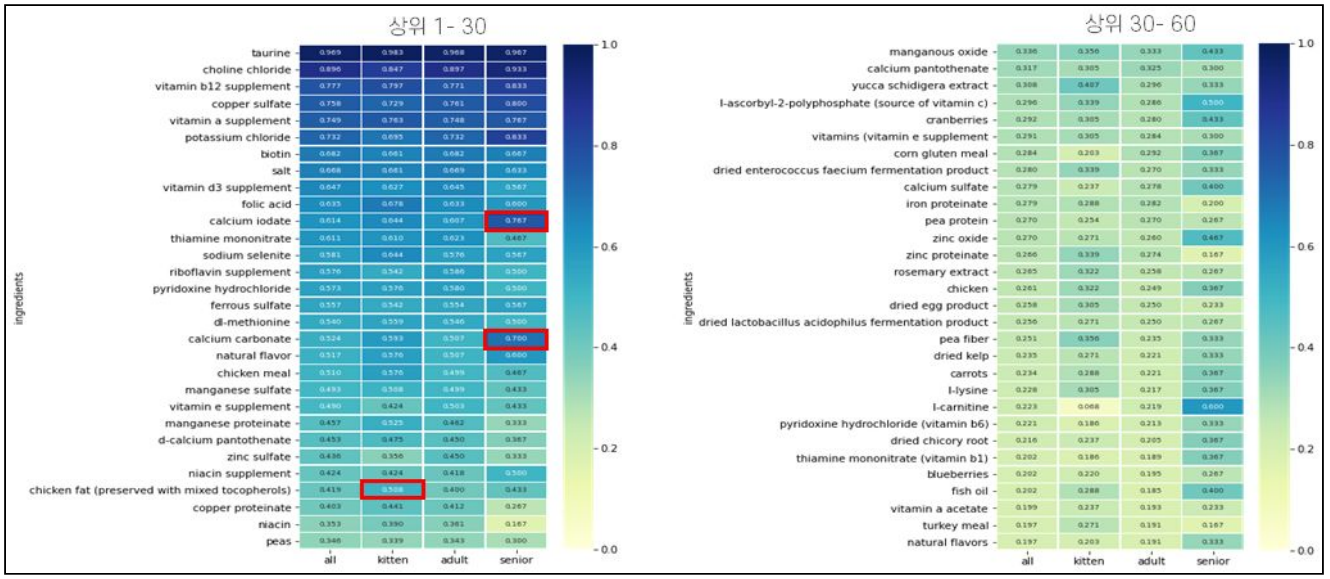
4) 연령대별로 특이한 원료를 분석

ingredients	all	all_rate	kitten	kitten_rate	adult	adult_rate	senior	senior_rate
fish oil (source of ara-arachi	3	0.00519	3	0.050847	0	0	0	0
fish oil (source of dha)	2	0.00346	2	0.033898	0	0	0	0
potassium chloride. a453118.	1	0.00173	1	0.016949	0	0	0	0
dried bacillus coagulans fern	1	0.00173	1	0.016949	0	0	0	0
potassium chloride. j453018.	1	0.00173	1	0.016949	0	0	0	0
organic sweet potato	1	0.00173	1	0.016949	0	0	0	0
d-pantothenic acid	1	0.00173	1	0.016949	0	0	0	0
powdered psyllium seed hus	1	0.00173	1	0.016949	0	0	0	0
salmon oil (preserved with m	1	0.00173	1	0.016949	0	0	0	0
th iamine mononitrate	1	0.00173	1	0.016949	0	0	0	0
menhaden herring meal	1	0.00173	1	0.016949	0	0	0	0
broccoli.	1	0.00173	1	0.016949	0	0	0	0
sesame oil (preserved with m	1	0.00173	1	0.016949	0	0	0	0
salmon oil (source of omega	1	0.00173	1	0.016949	0	0	0	0
dried bacillus coagulans fern	1	0.00173	1	0.016949	0	0	0	0

→ 연령대별로 특이한 원료를 분석한 결과, 그 수가 너무 적어 연령대별 특징적인 원료가 아닌 사료 하나에만 들어있는 원료로 파악했다.

연령대 그룹에서 공통으로 나타나는 원료들 중에서 연령대별로 빈도 수 차이가 있는 것을 분석했다..

5) 연령대별로 빈도수 차이 비교



→ 특정 연령대에 눈에 띄게 많이 나타나는 원료들을 확인할 수 있었다.

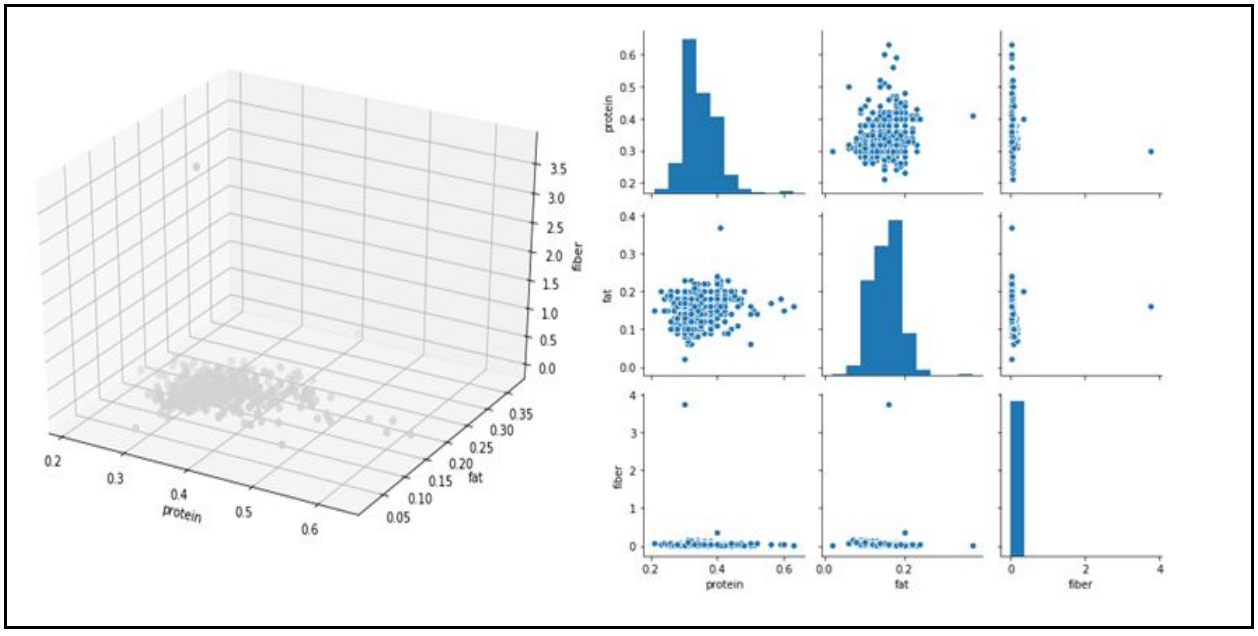
소비자가 원료를 보고 어떤 원료가 반려묘에게 필요한지 알기 어렵기 때문에 연령대별 빈도수가 높은 원료들을 토대로 추천 시스템에 활용할 수 있다. 이를 통해 명칭만 길게 나열되어 있어 연령대 별로 어떤 원료가 필요한지 쉽게 파악하기 어려웠던 기존의 문제점도 해결할 수 있다.

2.2 영양 성분 함량별 분석

영양 성분 함량별로 나뉜 cluster 간 원료를 wordcount하여 원료의 패턴 차이가 존재하는지 확인한다.

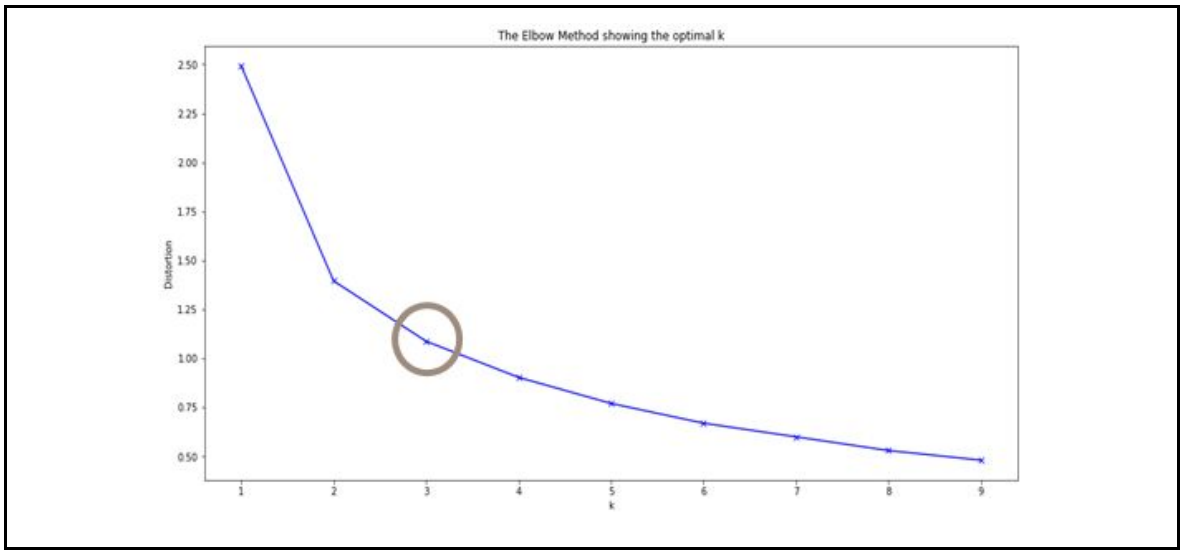
protein, fat, fiber 기준 K-means Clustering

1) 산점도 파악 후 영양 성분 간 상관관계 파악



→ 영양성분 간에는 유의미한 상관관계는 없는 것으로 확인됨.

2) K-means를 진행하기에 앞서 K 개수를 결정하기 위한 Elbow method algorithm 적용



→ Elbow point=3 으로 K=3으로 K-means Clustering 진행

3) 영양 성분 함량 기준으로 Clustering 진행

```
def kmean(n,k):
    d = n.shape[1]
    c = np.random.choice(range(len(n)),k)
    print('n:',len(n))
    c = n[c].astype('float64')
    print(c)
    init = True
    while init == True:
        cluster = np.zeros((len(n),k))

        for i in range(len(n)):
            temp = np.zeros((k))


            for j in range(k):
                g_temp = 0
                for g in range(d):
                    g_temp += (c[j][g] - n[i][g])**2

                temp[j] = g_temp**0.5

            cluster[i][np.argmin(temp)] = 1
```

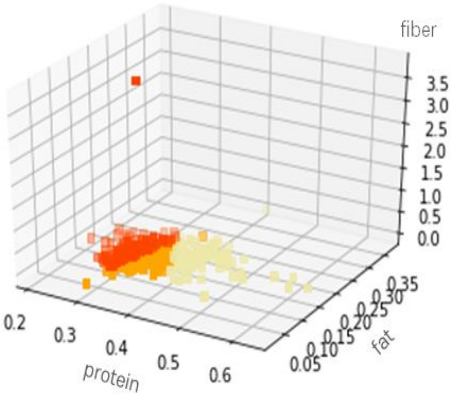
```
    cnt = 0
    for p in range(k):

        clu_index = np.where(cluster[:,p]==1)
        clu_vec = n[clu_index]
        print("np.sum",np.sum(clu_vec, axis=0))
        print(len(clu_vec))
        k_to_c = np.sum(clu_vec, axis=0)/len(clu_vec)
        point_c = 0
        for g in range(d):
            point_c += (c[p][g] - k_to_c[g])**2


        if point_c**0.5 < 3:
            cnt += 1
            if cnt == k:
                init = False
                break

        c[p] = k_to_c

print(c)
return cluster, c
```



4) Clustering 결과 Outlier 발견 Isolation Forest를 활용한 Outlier detection

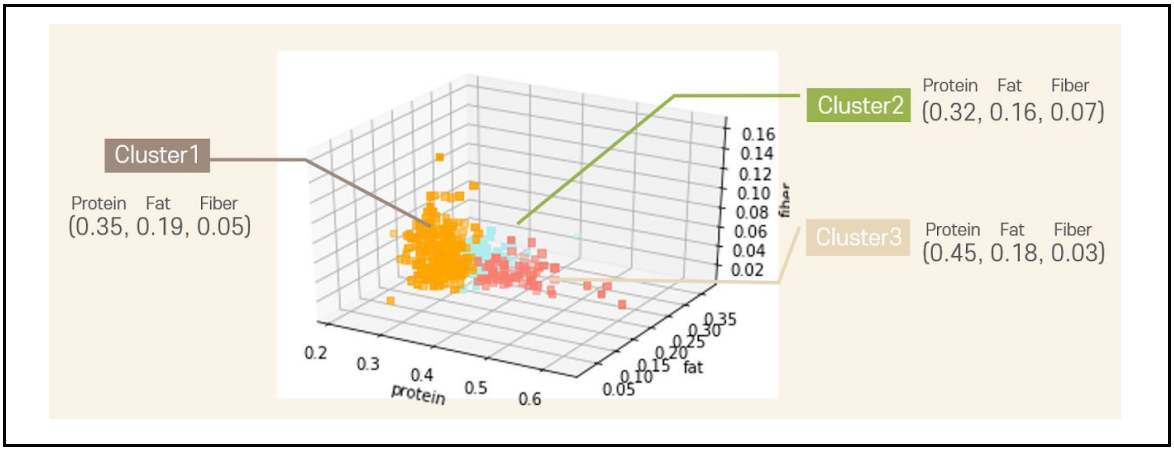
```
[21, 107, 170, 238, 249, 495]
 1    572
-1      6
Name: anomaly, dtype: int64
```

	protein	fat	fiber	anomaly
0	0.31	0.11	0.04	1
1	0.30	0.10	0.06	1
2	0.32	0.12	0.03	1
3	0.28	0.10	0.03	1
4	0.33	0.13	0.04	1
...	...	...	...	...
107	0.32	0.09	0.16	-1
170	0.50	0.06	0.08	-1
238	0.31	0.07	0.16	-1
249	0.38	0.11	0.13	-1
495	0.30	0.16	3.75	-1

	protein	fat	fiber	anomaly
0	0.31	0.11	0.04	1
1	0.30	0.10	0.06	1
2	0.32	0.12	0.03	1
3	0.28	0.10	0.03	1
4	0.33	0.13	0.04	1
...	...	...	...	...
573	0.33	0.16	0.04	1
574	0.52	0.14	0.06	1
575	0.31	0.10	0.09	1
576	0.35	0.19	0.04	1
577	0.36	0.22	0.04	1

→ Outlier Detection 결과 Outlier 6개 발견, Outlier를 제거하고 다시 K-means 시행

5) Outlier 제거 후 K-means 재실시



위의 클러스터링 결과로 클러스터별 word count를 진행해 특이적 원료를 찾는다.

Cluster 원료 패턴 wordcount

1) cluster 구분

```
#군집분석
clu,c = kmean(data3,3) #clu(데이터그룹), c(클러스터 위치)

n: 572
[[0.35 0.19 0.05]
 [0.32 0.16 0.07]
 [0.45 0.18 0.03]]
```

cluster 구분: 3개의 영양성분 상대 비교

cluster1: fat이 높은 cluster

cluster2: fiber가 높은 cluster

cluster3: protein이 높은 cluster

→ cluster별 원료 word counting 진행

2) cluster별 빈도수 상위 10개 원료 추출 결과

cluster1		cluster2		cluster3	
ingredients	frequency	ingredients	frequency	ingredients	frequency
vitamin	905	wheat	905	flour	658
sulfate	394	eggs	658	cartilage	360
calcium	360	tallow	632	papayas	336
chicken	296	sunflower	394	cod	296
chloride	258	marigold	336	tea	293
fermentation	255	chickpeas	296	grain	255
proteinate	247	prebiotic	293	garbanzo	219
zinc	215	chloride	258	pollock	202

word count 분석 결과, 상위에 taurine, vitamin, mineral 등이 등장하는 것을 확인하였다. 해당 원료들을 찾아보니 AAFCO에서 비타민과 미네랄을 필수 영양성분으로 지정하여 원료에 표시하도록 권고하고 있었다. 'chewy'에서는 원료 정보에 포함되어 있어서 원료로 추출을 했는데 해당 부분은 원료가 아닌 영양성분이기 때문에 이를 제외한 원료들만으로 word count를 다시 진행하였다.

3) 미네랄, 비타민등 원료표시를 제외한 단어 빈도수 확인

High - Fat

cluster1

ingredients	frequency
chicken	296
pea	180
fish	95
rosemary	86
potatoes	76
rice	74
egg	73
soybean	70
chicory	69
salmon	69
cranberries	66
turkey	64
corn	57

High - Fiber

cluster2

ingredients	frequency
wheat	905
eggs	658
sunflower	394
chickpeas	296
fish	255
fatty	219
crab	211
shells	204
rabbit	128
avocado	118
duck	99
caramel	94
flakes	92

High - Protein

cluster3

ingredients	frequency
flour	658
tea	293
grain	255
pork	132
tuna	131
potatoes	128
wheat	128
hearts	107
peas	100
oat	96
celery	94
grass	92
soybean	75

Cluster1 (High – Fat)

Cluster2 (High – Fiber)

Cluster3 (High – Protein)

→ 클러스터링 결과를 토대로 fat, fiber, protein 조절이 필요하면 해당 원료를 제외하고 사료를 추천해줄 수 있다.

소비자가 사료를 구매할 때 사료 판매 사이트에서 제공하는 분류에만 의존하지 않도록 영양성분이 어떤 원료와 관련있는지 보다 직관적이고 쉽게 알려줄 수 있다. 추가로 차원을 늘려 다른 영양성분 분석도 적용가능하다.

2.3 기타 분석 내용

1) 리뷰 분석 - word count

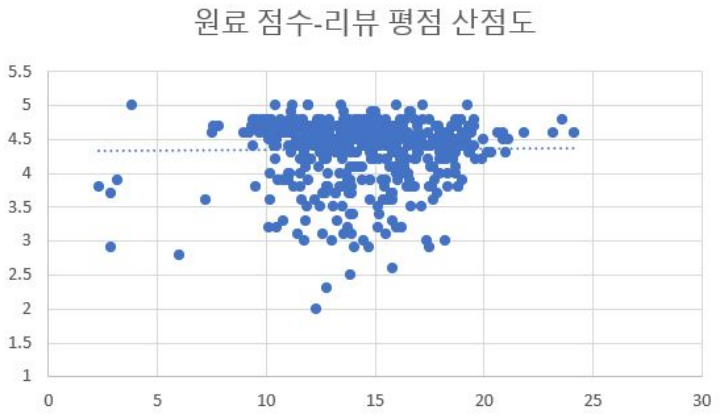
meow mix 사료로 리뷰 내용 텍스트 마이닝을 진행하였으나 eat, cat, food, 브랜드명, 사이트명 같은 단어들이 많이 카운팅 되어 의미 있는 정보를 찾기는 어렵다고 판단하였다. 아래는 'Meow Mix Original Choice Dry Cat Food' 사료의 리뷰를 word counting한 결과이다.

```
'cat': 1.0,
'food': 0.6111111111111112,
'love': 0.5,
'Meow Mix': 0.5,
'feed': 0.4444444444444444,
'eat': 0.3888888888888889,
'good': 0.3888888888888889,
'liked': 0.3333333333333333,
'Chewy': 0.3333333333333333,
'understand': 0.3333333333333333,
'tried': 0.2777777777777778,
'buy': 0.2777777777777778,
'feral cats': 0.2777777777777778,
'owner': 0.2222222222222222,
'first': 0.2222222222222222,
'Meow': 0.2222222222222222,
'smart': 0.2222222222222222,
'readily': 0.2222222222222222,
'BAG': 0.2222222222222222,
```

2) 원료점수 - 리뷰 평점 간의 상관관계 분석

많이 사용하는 원료와 리뷰 평점 간의 연관성이 있는지 분석해 해보았다. 연령대별 원료 등장 빈도로 사료별 원료 점수를 계산하였고 이를 바탕으로 원료점수와 리뷰평점의 산점도와 상관관계를 찾아보았다. 하지만 아래와 같이 상관관계를 파악할 수 없었다. 원료를 리뷰와의 연관성을 찾을 수 있는 객관적인 척도로 사용하기에는 무리가 있는 것 같다.

	score	ingredient	review_sta	review_nu
Meow Mix Original Choice	11.15779	39	4.7	960
Cat Chow Indoor Dry Cat F	10.57594	42	4.7	1000
Cat Chow Complete Dry Ca	10.05325	38	4.8	846
Kit & Kaboodle Dry Cat Fo	8.994083	36	4.6	669
Royal Canin Veterinary Die	14.28994	44	4.7	1864
Kitten Chow Nurture Dry C	9.372881	34	4.8	741
Cat Chow Naturals Origina	9.877712	36	4.7	828
American Journey Turkey &	20.90927	55	4.6	1465
9 Lives Daily Essentials with	12.71992	40	4.6	558
Iams ProActive Health Indc	10.17554	39	4.7	1099
Meow Mix Tender Centers	14.46154	48	4.7	497
Hill's Prescription Diet c/d I	13.83826	40	4.8	1219
Purina ONE Sensitive Skin	7.859961	39	4.7	675
Purina ONE Tender Selects	9.676529	43	4.7	800
Royal Canin Veterinary Die	14.85207	47	4.9	828

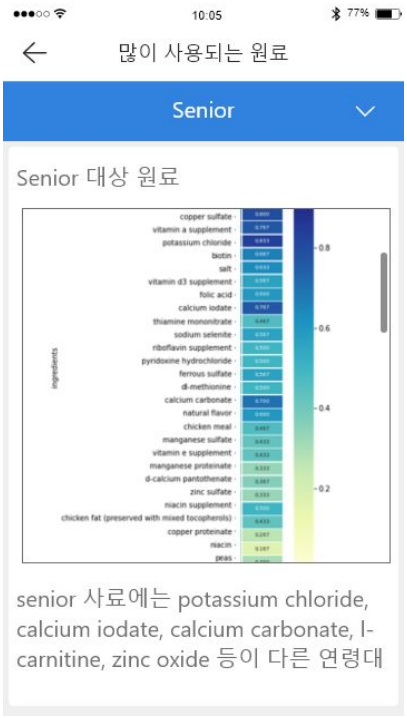


3. 어플리케이션 활용 시나리오

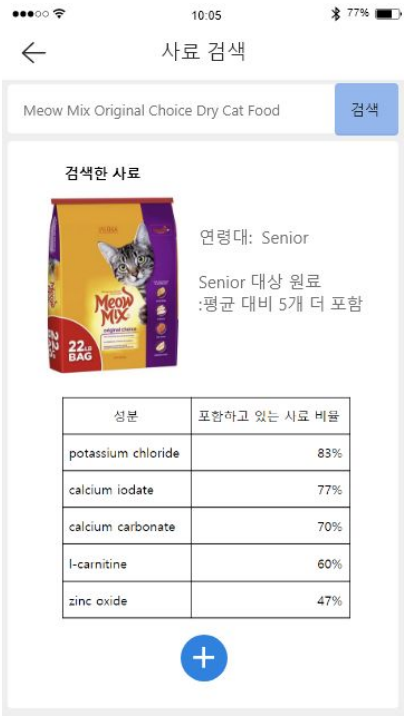
반려동물 사료를 구매할 때 어떤 원료가 많이 사용되고 필요한지 자세히 알지 못하는 소비자가 많다. 그래서 본 프로젝트를 통해 원료를 따져보고 사료를 구매하고 싶은 소비자를 위한 서비스를 제공하고자 한다

1) 연령대별 워드 카운트

어떤 사료가 반려묘에게 더 좋을지 찾고 싶을 때, 궁금한 사료를 입력하면 영양 성분 함량을 비교해주고, 연령대별로 원료 등장 빈도수에 가중치를 두어 반려묘의 연령대에 적합한 사료인지 알려준다.



[화면1] 연령대별 많이 사용되는 원료

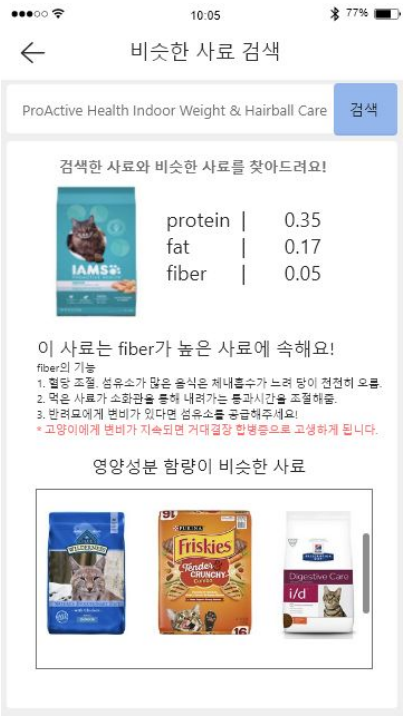


[화면2] 사료명 검색 결과

왼쪽은 제공하고자하는 서비스를 어플에 적용한 예시이다. 소비자는 반려묘 사료가 연령대별로 어떠한 원료들을 많이 포함하고 있는지 확인할 수 있다. 예를 들어, [화면1]과 같이 Senior용 사료에 많이 들어있는 원료에는 어떤 것이 있는지 원료 빈도 수 그래프와 설명을 확인할 수 있다. 그리고 사료명을 검색하면 그 사료의 연령대와 해당 연령대에서 많이 등장하는 원료를 포함하고 있는지 상대적으로 비교한 결과를 알 수 있다. [화면2]는 사료 검색 결과이다.

2) 클러스터링

[화면3] 조절 성분 검색 결과



[화면4] 비슷한 사료 검색 결과

fat, protein, fiber 중 반려묘에게 조절이 필요한 성분을 검색하면 fat이 높은 클러스터 특성을 반영하여 검색 결과를 보여준다. [화면3]과 같이 fat을 검색하면 피해야하는 원료로 fat이 높은 클러스터의 원료 word count 결과가 나타난다. 또한 그 원료들을 포함하고 있는 사료를 조심해야하는 사료로 알려준다.

클러스터링 결과로 비슷한 사료 검색에도 활용할 수 있는데 사료를 검색하면 사료의 영양 성분 함량과 특성을 알려주고 영양 성분 함량이 비슷한 사료를 추천해준다.

4. 한계 및 개선 방안

본 프로젝트의 한계점과 개선 방안을 살펴보고 마무리하자면, 우선 원료 분석의 통일성 문제이다. 클러스터 분석 과정에서 원료로 추출된 비타민과 미네랄 등이 영양성분에 해당한다는 것을 파악하게 되었고, 이를 연령대별 원료 분석에 반영하지 못했다. 이 부분은 후에 원료 분석 과정을 통일하여 개선할 수 있다. 두 번째는 center를 기준으로 각 클러스터의 특징을 잡아 상대적으로 비교했다는 점이다. protein, fat, fiber 만으로 사료의 영양성분을 일반화하기에는 무리가 있다. 이는 차원을 추가해 다양한 영양성분을 함께 고려한다면, 보다 정확한 지표로 활용할 수 있을 것이다. 마지막으로, 리뷰와 평점 데이터에서 유의미한 결과를 얻지 못하였는데, 묘종별 사료 선호도를 추가한다면 묘종별 사료추천과 함께 묘종별 리뷰 비교 등으로 확장해 볼 수 있을 것이다.

참고문헌

[1] chewy, 반려동물 용품 판매 사이트 https://www.chewy.com/

[2] 이뉴스투데이 http://www.enewstoday.co.kr

[3] 반려동물 트렌드 리포트 2020. opensurvey https://blog.opensurvey.co.kr/article/companionanimal-2020-2/

[4] 국내 반려동물 식품 및 용품 시장현황 분석 연구, 한국콘텐츠학회논문지, http://asq.kr/Xs7WQPDeTtTg6

[5]영양성분 프로파일링 기반 사료추천 알고리듬, 한남대학교, 송희석. https://url.kr/m4hnub