# Capstone project 1: Consolidated Report

Project Title:
Pet Product Subcategorization by Review Analysis

**< Table of Contents >**



## 1. Problem Statement

### 1.1. Problem

E-commerce companies set up categories for their products; for example, Clothing, Beauty, Books, Pet Supplies, etc. However, if the number of products in a category has been growing, they might want to classify the products into subcategories for several reasons. Because the number of products would be large, it could be difficult to categorize them one by one.

Nowadays, most e-commerce websites have reviews, which are written by customers and which help future customers decide whether they buy. On the other hand, reviews also provide much information to an e-commerce company; what customers liked or disliked, what they wanted, how they used it, and so on. In other words, reviews include information to categorize a product into subcategories. A product title usually has a lot of information about the product, but sometimes reviews have unpredictable information from the title. For example, when I was looking for a new ‘cat litter’, I found that some specific kinds of ‘cat litter’ were popular as rabbit litter from the reviews.

Here, I had chosen Pet Supplies as a category reclassified into subcategories. The products in Pet Supplies were subcategorized into small categories by the review analysis.

### 1.2. Client

The primary clients would be e-commerce companies that would like to use reviews to subcategorize their products. Their purposes could be to improve analysis of trends and customer needs to a specific field, and/or to increase customer satisfaction by easy access to a product they want. Also, this technique can be used by a manufacturer to extract a specific kind of products from a big category to analyze their competitions.

## 2. Data preprocessing

### 2.1. Data Source

The animal product review dataset was acquired from AWS. The data contained information about the marketplace, product ID, product name, product category, star rating, review, date, etc. In this project, I focused on the data collected during 2014-2015 in the US.

- Amazon Pet Supplies reviews in the US (gz file) from AWS (1995 – 2015)
  https://s3.amazonaws.com/amazon-reviews-pds/tsv/amazon_reviews_us_Pet_Products_v1_00.tsv.gz

### 2.2. Data wrangling

The purpose of this section was to make the data ready for clustering. The necessary data were extracted, and the reviews were divided into tokens in the following steps.

Extract the necessary rows and columns, and deal with the missing and duplicate data:

- Extract data collected in 2014 or 2015
- Remove columns having a single value
- Deal with missing values and duplicate data

Collect a moderate amount of fitted reviews:

- Remove short and long reviews
- Extract 10 reviews per product

Clean the tokens:

- Tokenization
- Retain only alphabets
- Remove stop words
- Stemming and lemmatization
- Retain only nouns
- Remove words appearing in less than 5 products

There were 1,705,229 observations (reviews) and 128,995 products in the data after extracting the data collected in 2014 or 2015. Missing values were dropped because the rate was low. There were 216 (0.0013%) observations that lacked a review body. 18 products (0.0014%) were removed from the data by removing the observations (missing data). There were 8,004 duplicate rows. They were purely removed from the data. In consequence, 1,697,225 reviews remained, but the number of products in the data did not change (128, 977 products).

Short reviews (less than 30 characters) were removed. If a review was longer than 281 characters (75th percentile),

the first 281 characters were extracted and the rest of the review was cut. As a result, 9,780 (7.6%) products were dropped. A short review would not have enough information to subcategorize the product, and the products having only short reviews would not be very active. A long review would have too much information, and it could make the model complicated.

Some products were very popular and had a lot of reviews. On the other hand, some products were not so popular and had a few reviews. This difference would increase in complexity of clustering if all of the reviews in the data were used. That was why the number of reviews per product was adjusted so that each product had the exact ten reviews. When a product had fewer reviews than 10, it was removed. When a product had more than 10 reviews, the latest reviews were selected for the product. As a result, 84% of products in the data were removed, and popular 20,403 products (having more than 10 reviews) remained. Then, the reviews to the same product were merged so that each product had one long review for the next step.

The next step was tokenization. The sentences were chopped into words. Then, words that would not be used to predict were handled in the following steps. First, non-alphabet words and stop words were removed. The remaining words were transformed into the base forms through stemming and lemmatization. Second, I determined that words other than nouns would not be necessary to achieve the goal, and removed non-noun words by using POS. Third, words appearing only in 5 or fewer documents were dropped because the words would not only have much impact but also increase the number of dimensions and make the modeling complicated. Finally, the words in the Top 200 appearance that were not characteristic were removed by hand.

## 3. Exploratory data analysis

### 3.1. Tokens

There were 6,937 unique words in the Dataset after preprocessing. The frequent words (the top 100) were shown in figure 1. Most of the frequent words looked important, and they included animal kind words (dog, cat, bird, etc.) and application type words (food, toy, treat, tank, etc.) These words were a significant clue to subcategorize products.

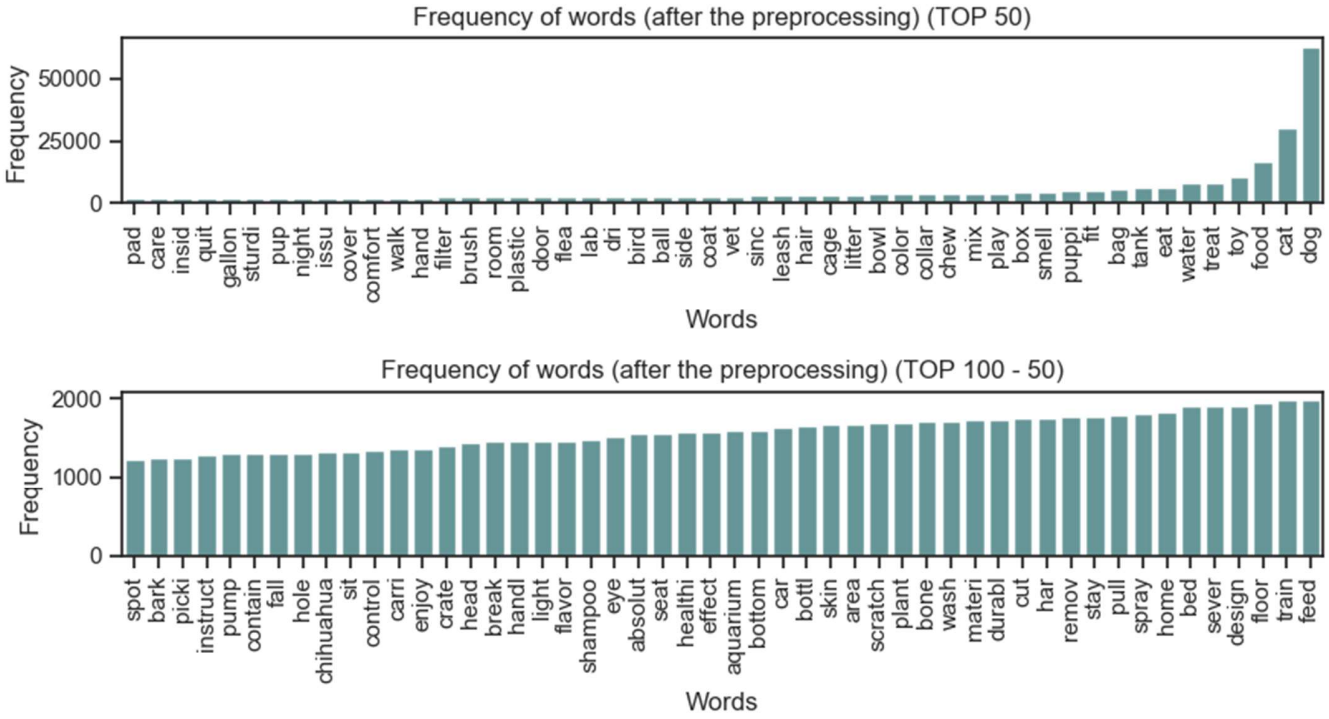



**Figure 1. Frequency of words (Top 100)**

Next, the number of tokens per product were counted. The results were shown in Figure 2. The minimum was 5 tokens, and the maximum was 126.

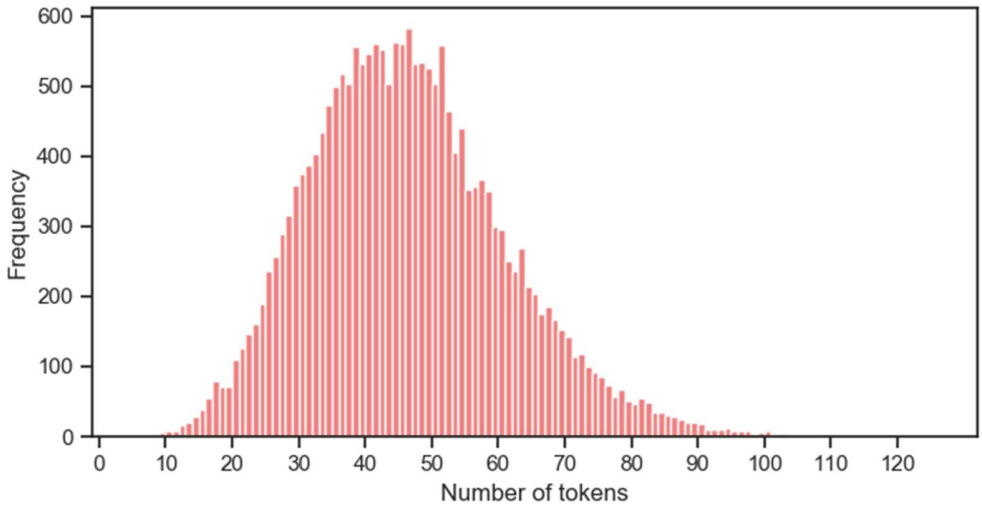



**Figure 2. The number of tokens per product**

### 3.2. The number of products relating to each animal kind

Several animals were found in the tokens above. Here, I counted the rough numbers of products relating to each specific animal. I picked up some animals: dogs, cats, birds, fish, and rabbits. The keywords I chose for each animal category were shown in Table 3. When a product had the dog keywords (such as ‘dog’), the product was counted as

a dog-related product.

**Table 3. The keywords for each animal category**

| Category | Keywords |
|---|---|
| **Dog** | dog, puppi, doggi, pup |
| **Cat** | cat, kitti, kitt, kitten |
| **Bird** | bird, chick |
| **Fish** | fish |
| **Rabbit** | rabbit, bunni |

One product could be counted as both dog-related and cat-related when the product had both tokens (such as 'dog' and 'cat'.) It might mean the product could be used for dogs and cats. Figure 3 showed the number of products relating to each animal.

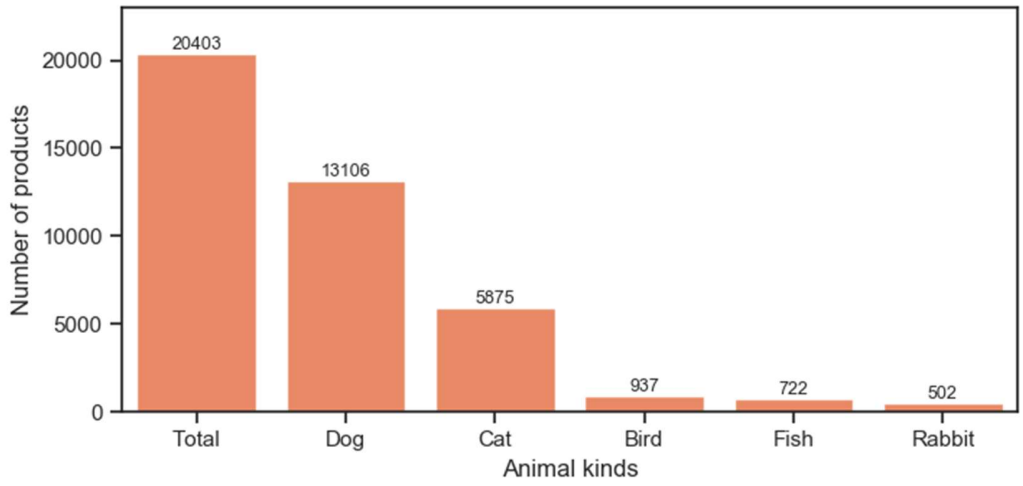



**Figure 3. The number of products relating to each animal kind**

Dog products and cat products were the largest two groups. Especially, 64% of the products were related to dogs. I decided to separate the data into three datasets (dog, cat, and other) to make the clustering efficient (see 3.5.).

### 3.3. The number of products relating to each application type

There were tokens such as food, toy, treat, tank, collar, cage, leash, bowl, etc. in the data. Here, I prepared 7 categories for application types. The keywords for each category were shown in Table 4. Each category was counted in the same manner as counting animal kinds. The result was shown in Figure 4.

This rough calculation was useful to estimate the number of clusters and the sample sizes in the clustering step. The Top 3 categories were Food, Treat, and Toy. However, the products were relatively spread in the seven categories, and 35% of the products were in the other categories. This would be because the keywords I picked up mainly fit

products for dogs and cats. There would be unique products that were not classified into them, especially products for fish.

**Table 4. The keywords for each application category**

| Categories | Keywords |
|---|---|
| **Toy** | toy, tunnel, ball, rope |
| **Food** | food, dri, wet |
| **Treat** | treat, snack, cooki |
| **Collar & Leash** | collar, leash |
| **Clothes** | shirt, coat, sweater, costum |
| **Cage** | cage, crate, carrier, kennel |
| **Toilet** | litter, pad |

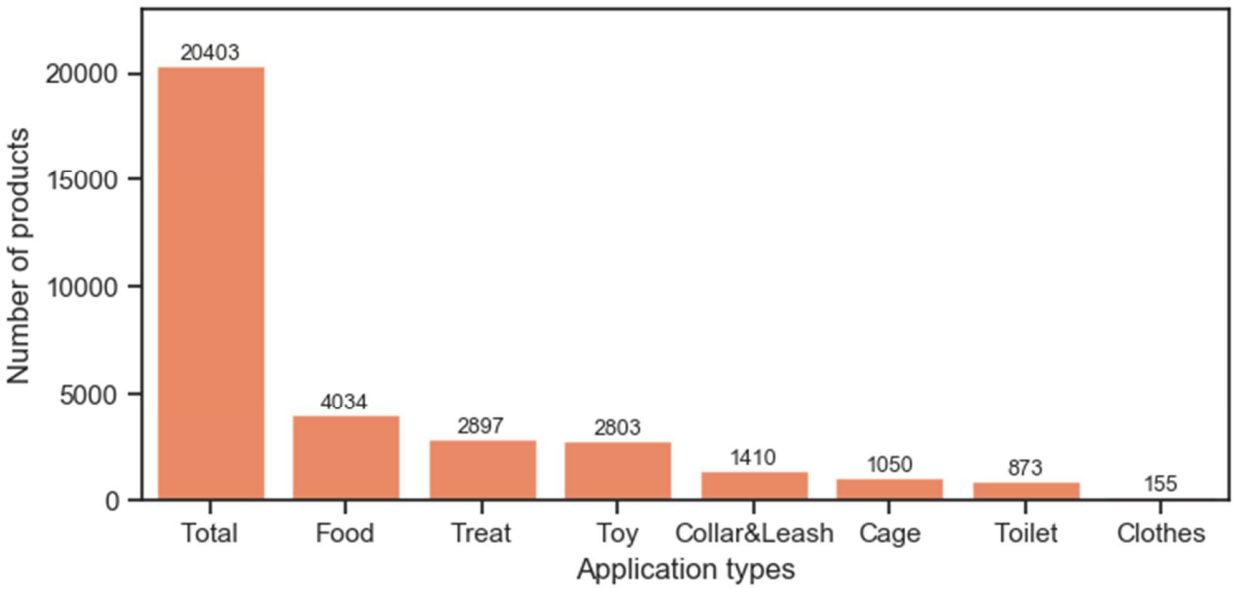



**Figure 4. The number of products relating to each application category**

### 3.4. The rough relationship visualization between an animal kind and application type

Here, Dog and Cat were picked up as animal categories, and Toy and Food as application categories here. Then, the relationships between them were estimated. This meant that a product was classified as a dog- and toy-related product if the product had tokens such as ‘dog’ and ‘toy’. The result was shown in Figure 5.

Food and Toy for dogs were the largest two categories on the chart, but many products were in Other for dogs or Other for the other animals. It was interesting to note that there were certain amounts of products in Toy & Food for dogs and cats. These could be a food dispenser ball or food puzzles, which a dog and cat could play with and eat from. These products could be potentially difficult to be categorized into one subcategory because they would have

both features; toy and food.

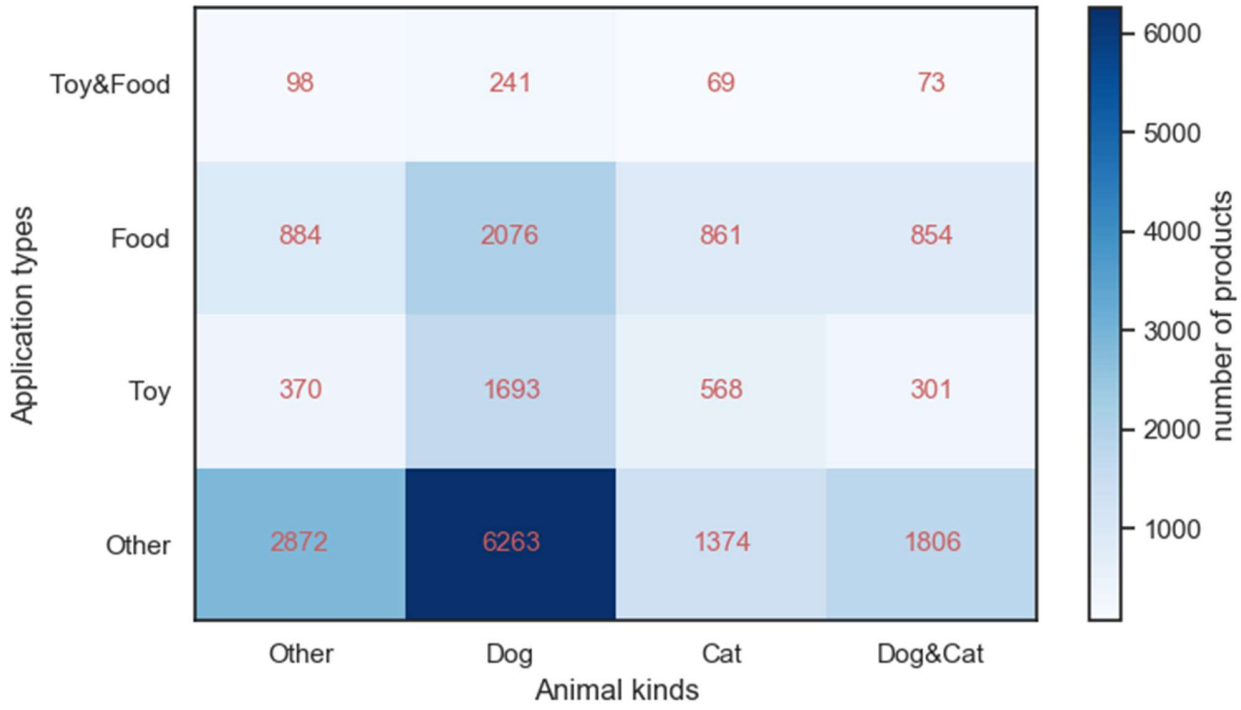



**Figure 5. The relationship between animal kinds and application types**

### 3.5. Divide the data into three datasets

There were 20,403 products in the dataset at this moment. It was still big, and it was expected that the subcategories themselves could be very different depending on animal kinds (3.3.). That was why the data were divided into three sets; dog, cat, and other. The numbers of the keywords for each animal category were counted to each product. If the dog keywords were the most frequent, the product was put into the dog dataset. So was the cat dataset. The products were classified into the 'other' dataset in the other case. Therefore, there were no duplicate data points between the three datasets. Each product was set to belong to the most likely dataset.

The numbers of products and unique tokens were shown in Table 5. The dog dataset had 11,916 products and was the biggest dataset. Each dataset had about 6,000 kinds of tokens.

**Table 5. Three datasets**

| Category | Total products | Total tokens (unique) |
|---|---|---|
| **Dog** | 11,916 | 6,682 |
| **Cat** | 4,099 | 6,001 |
| **Other** | 4,388 | 5,914 |

The summary statistics of the number of tokens were shown in Table 6. Some products had a few tokens, and some products had many tokens in all datasets. These differences could affect the performance of clustering. To adjust the differences, the matrixes were normalized before clustering.

**Table 6. The summary statistics of the number of tokens**

| Category | Min | 25 % | 50 % | 75 % | 90 % | Max | Mean | SD |
|---|---|---|---|---|---|---|---|---|
| **Dog** | 7 | 36 | 45 | 55 | 66 | 126 | 46.0 | 14.8 |
| **Cat** | 9 | 38 | 47 | 58 | 69 | 120 | 48.7 | 15.1 |
| **Other** | 5 | 32 | 42 | 52 | 63 | 98 | 43.1 | 14.7 |

## 4. Clustering

### 4.1. Vectorization

Count vectorizer was used to vectorize the tokens because the frequent words seemed useful to describe each product (Figure 1). Each dataset had about 6,000 kinds of tokens. Additionally, bigrams were used so that some frequent compound words could be recognized as one word. For example, 'dried food', 'training pad', and 'litter box'.

In the case of the dog dataset, if all of the single tokens and bigrams were used, the number of tokens was more than 10,000. I decided to use the top 5,000 tokens for clustering to control the number of tokens. Otherwise, too many dimensions would cause very long compute time and complications from the curse of dimensionality. In the case of the cat dataset and the ‘other’ dataset, I used a slightly different approach. The total number of single tokens and bigrams were not very big, and tokens appearing less than 6 times in the total documents were removed to control the number of tokens. As a result, 4,665 and 5,156 tokens were used for clustering respectively.

### 4.2. Hierarchical clustering

First, imagine how the result of this project would be used:

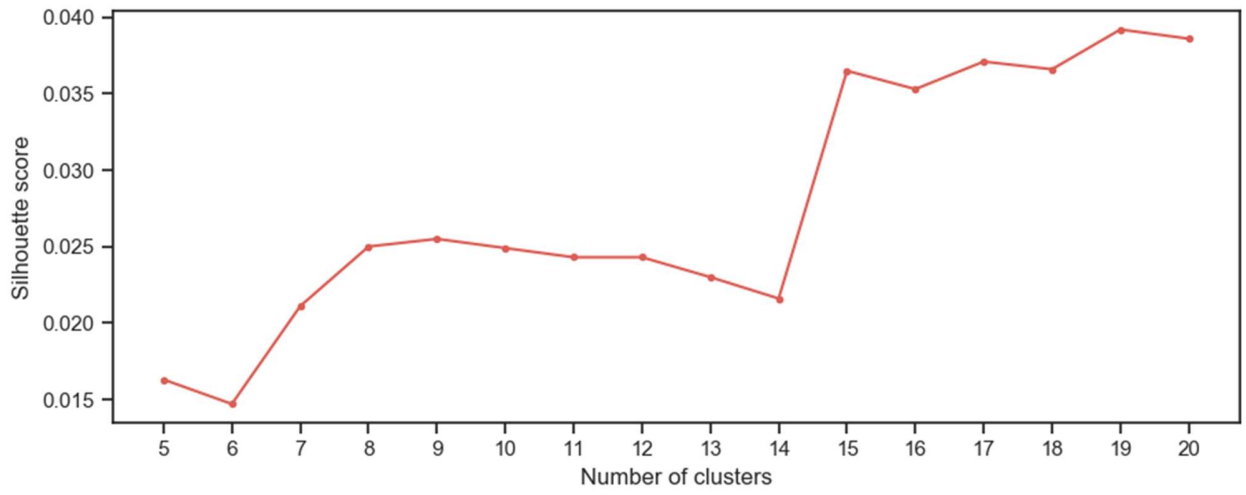
1. A customer visits the website of the e-commerce company.
2. Select 'Animal products' category
3. Select 'Cat' category
4. Select 'Grooming' category
5. Select 'Brush, Clipper' category
6. Browse products, find something interesting and buy it

As you can see, there are animal categories (we already had them; 'Dog', 'Cat', 'Other'), and under the animal categories there are some big categories (e.g. 'Grooming', 'Food', 'Toy'), and some small categories (e.g. 'Brush, Clipper', ‘Cat tree', 'Collar') under the big categories. Here, I chose hierarchical agglomerative clustering as an algorithm to achieve this. Hierarchical agglomerative clustering is a method of cluster analysis, which is one of unsupervised learning. Cluster analysis is generally used to segment data into some groups without any pre-labels. The feature of hierarchical agglomerative clustering is to build nested clusters by merging the clusters (at the start point, they are individual samples) until becoming one cluster successively. The hierarchy is represented as a

dendrogram. It was suitable for this project. The one weak point of this clustering is the time complexity; that is, it's slow. The three datasets of this project (dog, cat, other) had 4,000 to 12,000 data points respectively. It took time, but it was acceptable. When the function of hierarchical agglomerative clustering was used, there were two important parameters; affinity and linkage. The data were about texts this time, and Cosine similarity was chosen as the affinity. Average linkage was selected as the linkage because it was expected to have some big clusters and some smaller clusters.

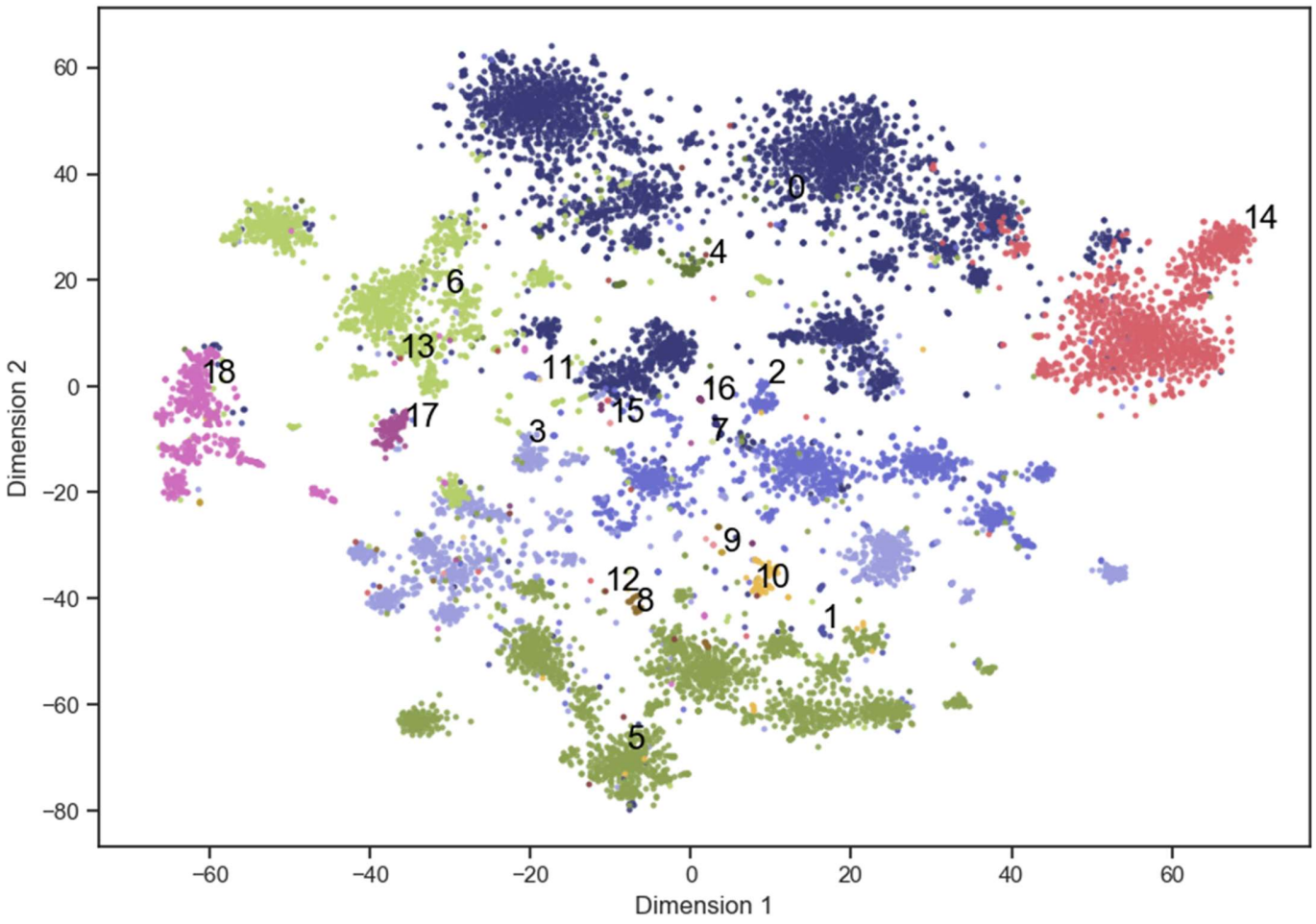
First, I set the number of the big categories to less than 10 for each dataset. The silhouette scores from k (the number of clusters) = 5 to 20 were calculated, and the k having the highest score was chosen. When the k was used for clustering, there were some big clusters and several mini clusters. The big clusters were considered as the big categories, and the mini clusters were count as the ‘other’ group. The products in the big categories were divided into smaller clusters again in the same way to get the small categories.

For example, the dog dataset was divided into 19 clusters (figure 6). According to the Silhouette plot, the t-SNE image with the labels (figure 7), the frequent words, and browsing the product titles in each cluster, cluster_0, 2, 3, 5, 6, and 14 had individual features, and I evaluated that they were appropriate as the big categories respectively. Cluster_18 was similar to cluster_6, and they were merged. Each big category was given the category name by browsing the product titles. The other mini clusters had fewer products. They were count as the 'other' group. However, some of them also had a unique character. They were also labeled, and the labels were used as the small categories of the ‘other’ group (table 7). Here, I found outliers in the mini clusters. All products in cluster_15 were about an aquarium. These mini clusters didn’t join up with the big clusters by the upper course of the dendrogram, and that meant the mini clusters were not very similar to the main clusters. As a result, some outliers could be detected in this way.




* The dog dataset. K = 19 was chosen in this case.

**Figure 6. The silhouette scores for the various number of clusters (the first split)**




* The dog dataset. k = 19

**Figure 7. The t-SNE plot colored by the labels (the first split)**

**Table 7. The category names (the first split)**

| Big categories | | "Other" | |
|---|---|---|---|
| **Cluster No.** | **Cluster label** | **Cluster No.** | **Cluster label** |
| 0 | food, treat, treatment | 4 | calming |
| 2 | bed, crate, gate | 7 | monitoring |
| 3 | clothes | 8 | memorial |
| 5 | collar, leash | 10 | tie out |
| 6, 18 | body care, cleaning | 15 | aquarium |
| 14 | toy | 17 | eye care |
| - | - | 1, 9, 11, 12, 13, 16 | other |

* The dog dataset.

Then, the big clusters (cluster_0, 2, 3, 5, 6, 14, and 18) were extracted and divided into 35 smaller clusters. They were named in the same manner as the first sprit. As a result, each of the dog products had a big category label and a small category label.

The cat dataset and the ‘other’ dataset were treated in the same way as the dog dataset. Then, all of the products and labels were merged into one table (table 8.) Each row represented one product with the product id, title, animal category, big category, and small category. This table was the end objective!

**Table 8. The product table with the category labels**

| | product_id | product_title | animal | big_category | small_category |
|---|---|---|---|---|---|
| 0 | 70064 | Perfect Pet Soft Flap Cat Door with Telescoping FramePerfect Pet Soft Flap Cat Door with Telesco... | cat | door, cage, carrier, bed | door, cage |
| 1 | 119780 | ARK Naturals PRODUCTS for PETS 326066 4-Ounce Breath-Less Chewable Brushless Toothpaste, MiniARK... | dog | food, treat, treatment | treat |
| 2 | 202371 | Stella & Chewy's Freeze Dried Dog Food for Adult Dogs, Chicken Patties, 15 Ounce Bag - 2 PackSte... | dog | food, treat, treatment | food, bowl |
| 3 | 291967 | Premium Deshedding Brush for Dogs and Cats with Medium to Long Hair \| Veterinary Approved \| Rugg... | dog | body care, cleaning | brush, clipper |
| 4 | 490904 | Remington Coastal Pet R0206 GRN06 Rope Leash, 72-Inch, GreenRemington Coastal Pet R0206 GRN06 Ro... | dog | collar, leash | leash, harness |
| 5 | 593896 | Pet Food Can Covers Lids Set of 3Pet Food Can Covers Lids Set of 3Pet Food Can Covers Lids Set o... | cat | food, treatment | food, treat, water |
| 6 | 674575 | Scotch Pet Hair Roller 839RScotch Pet Hair Roller 839RScotch Pet Hair Roller 839RScotch Pet Hair... | other | other | brush, comb, clipper |
| 7 | 690871 | Petco Brooklyn 55 Gallon Metal Tank StandPetco Brooklyn 55 Gallon Metal Tank StandPetco Brooklyn... | other | fish, reptile | other |
| 8 | 798322 | Pet Dog Puppy Nonslip Canvas Sport Shoes Sneaker Boots Rubber Sole Size 5 Blue by MallofusaPet D... | dog | clothes | shoes |
| 9 | 800175 | Purina Pro Plan Focus Large Breed Formula Dry Dog FoodPurina Pro Plan Focus Large Breed Formula ... | dog | food, treat, treatment | food, bowl |

* The first 10 products. There were 20,403 rows in the actual table.

## 5. Conclusion

20,403 pet products were subcategorized in total. Each product had three labels; an animal category, a big category, and a small category. The animal category was composed of three classes; dog, cat, and other. Each animal category had several big categories and small categories under the big categories. As a result of the subcategorization, the products were classified into 73 kinds of groups (table 9). This table showed what the popular categories were. For example, 'treat' category in 'food, treat, treatment' big category of the dog products had almost the same number of products in 'food, bowl' category. Also, if you find something interesting in one of the specific categories in this table, you can extract the reviews of the products, and analyze them closely.

The result of this project can be used for product classifications on the website of an e-commerce company and for extracting a specific group of products to analyze them closely; for example, the variety, popular products in the group, consumer needs, and others.

**Table 9. The category names and the number of products**

| animal | big_category | small_category | count |
|---|---|---|---|
| dog | bed, crate, gate | bed | 424 |
| | | door, gate, crate | 475 |
| | | other | 26 |
| | | step | 100 |
| | | tie out | 70 |
| | body care, cleaning | brush, clipper | 559 |
| | | dryer, towel | 47 |
| | | ear cleaner | 93 |
| | | eye care | 116 |
| | | flea care | 275 |
| | | odor, stain, shampoo | 857 |
| | clothes | costume | 672 |
| | | other | 38 |
| | | shoes | 133 |
| | collar, leash | ID tag | 126 |
| | | collar | 1039 |
| | | leash, harness | 803 |
| | food, treat, treatment | food, bowl | 1431 |
| | | oral care | 78 |
| | | other | 10 |
| | | treat | 1494 |
| | | treatment, supplement | 447 |
| | other | calming | 76 |
| | | car seat, cover | 279 |
| | | memorial | 27 |
| | | monitoring | 10 |
| | | other | 91 |
| | | stroller | 74 |
| | | training | 97 |
| | | training pad | 264 |
| | | waste bag, carrier | 436 |
| | toy | toy | 1242 |

| animal | big_category | small_category | count |
|---|---|---|---|
| cat | collar, leash | collar | 114 |
| | | harness, leash | 28 |
| | door, cage, carrier, bed | bed | 193 |
| | | carrier, stroller | 107 |
| | | door, cage | 101 |
| | | other | 9 |
| | | perch, shelf | 26 |
| | | step | 10 |
| | | tent | 11 |
| | food, treatment | cat grass | 20 |
| | | food, treat, water | 1220 |
| | | pill, treatment | 202 |
| | | scale | 6 |
| | grooming | brush. comb | 174 |
| | | clipper | 37 |
| | | flea | 155 |
| | | nail cap, furniture protector | 49 |
| | litter, odor, stain | litter, litter box | 534 |
| | | odor, stain, shampoo | 191 |
| | other | calming | 19 |
| | | memorial | 19 |
| | | other | 15 |
| | toy, scratcher, cat tree | scratcher, cat tree | 307 |
| | | toy | 535 |

| animal | big_category | small_category | count |
|---|---|---|---|
| other | bird, rabbit, hamster | bird | 882 |
| | | chicken | 73 |
| | | chinchila | 18 |
| | | litter, bedding | 30 |
| | | other | 6 |
| | | rabbit, hamster | 283 |
| | fish, reptile | aquarium | 1937 |
| | | crustacean | 57 |
| | | food (fish, turtle) | 348 |
| | | food (reptile) | 89 |
| | | other | 16 |
| | | terrarium | 164 |
| | other | bowl | 29 |
| | | brush, comb, clipper | 81 |
| | | memorial, tag | 39 |
| | | other | 150 |
| | | treatment | 118 |
| | | waste bag | 35 |