Capstone project 1: In-Depth Analysis

Project Title:

Pet Product Subcategorization by Review Analysis

1. Clustering

1.1. Data

In the preprocessing steps, reviews of products having more than 10 reviews were extracted, cleaned, tokenized, and divided into three animal categories; cat, dog, and other (see Data Wrangling Report.) The numbers of products and unique tokens in each dataset were shown on table 1.

Table 1. Three datasets

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

1.2. Vectorization

Count vectorizer was used to vectorize the tokens because the frequent words looked useful to describe each product (see Milestone Report). 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 complication 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.

1.3. Hierarchical clustering

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

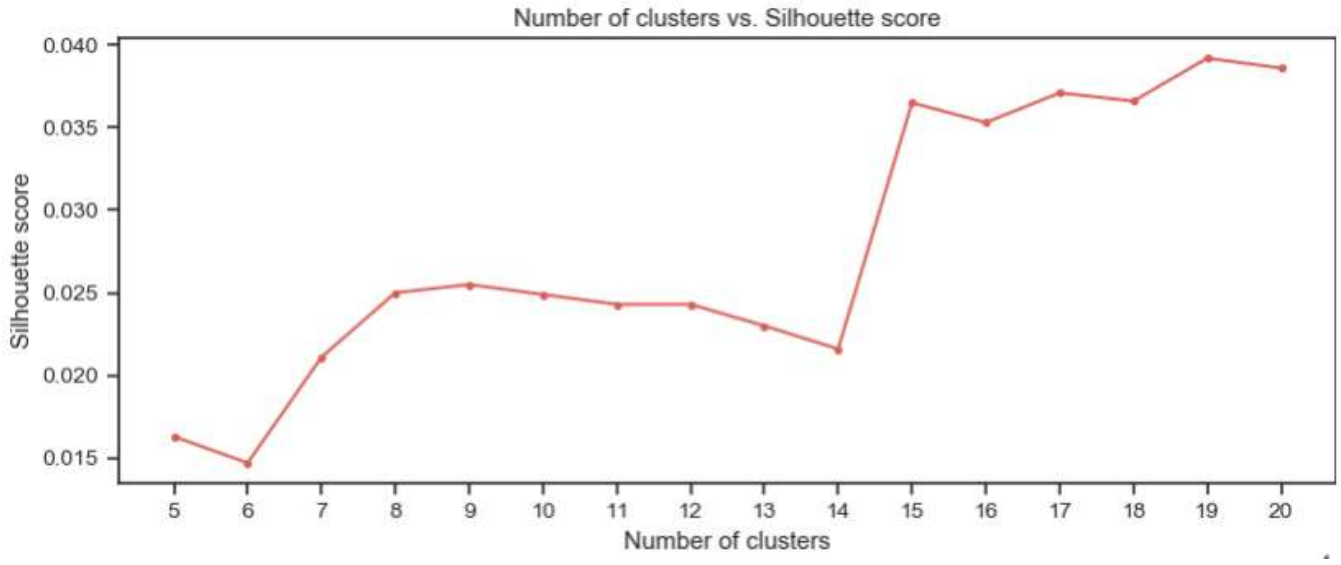
1. A customer visits the website of the e-commerce company.
2. Select 'Animal products' category
3. Select 'Cat' category
4. Select 'Grooming' category
5. Select 'Brush, Clipper' category
6. Browse products, find something interesting and buy it

As you 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', C'at 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 points 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 the purpose of 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 points 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 smaller clusters.

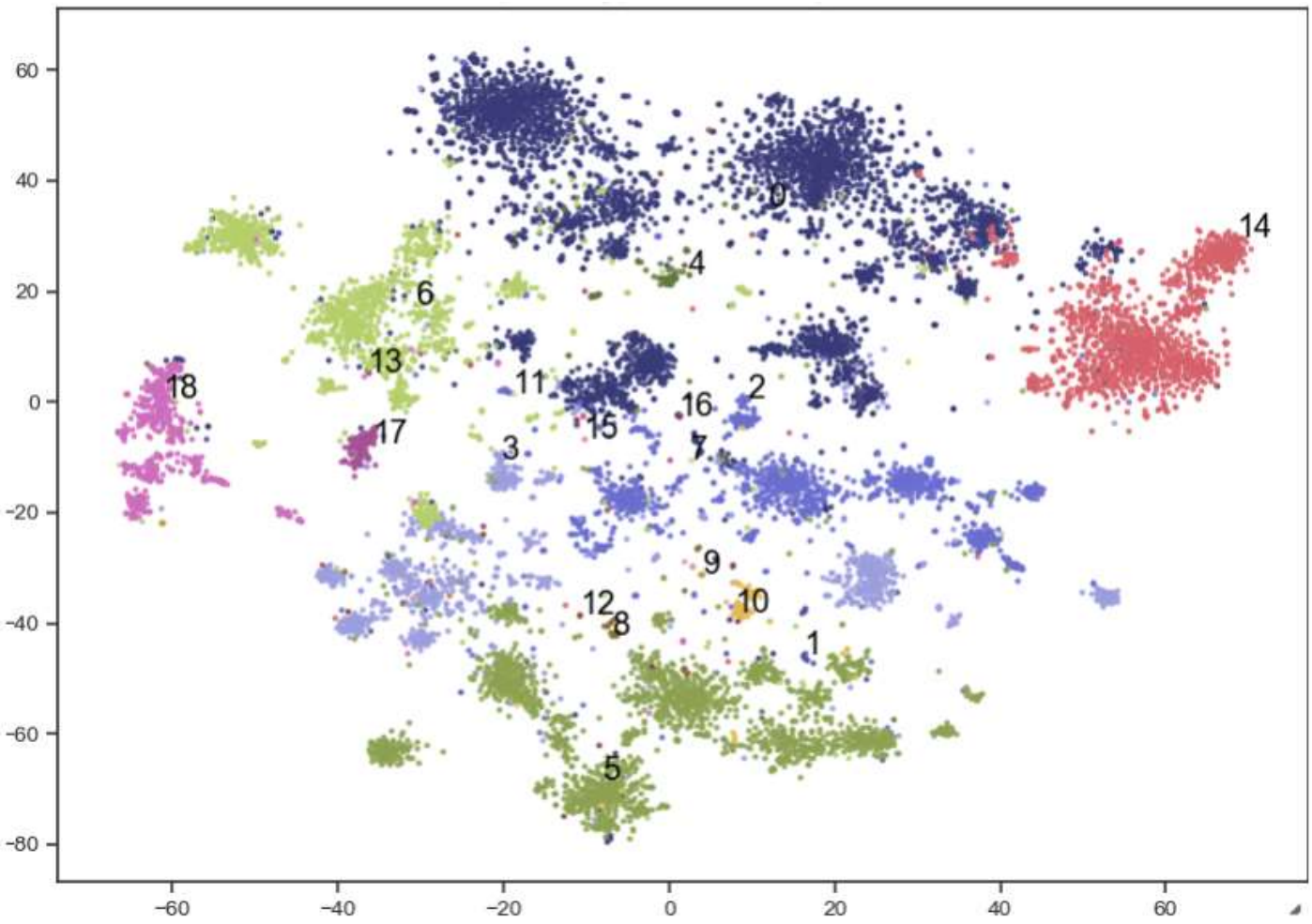
First, I set the number of the big categories to less than 10 for each dataset. The silhouette scores from k (the number of clusters) = 5 to 20 were calculated, and the k having the highest score was chosen. When the k was used for clustering, there were some big clusters and several mini clusters. The 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 1). According to the Silhouette plot, the t-SNE image with the labels (figure 2), 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 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 2). 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, 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 1. The silhouette scores for the various number of clusters (the first split)



* The dog dataset. k = 19

Figure 2. The t-SNE plot colored by the labels (the first split)

Table 2. 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 the products and labels were merged into one table (table 3.) Each row represented one product with the product id, title, animal category, big category, and small category. This table was the end objective!

Table 3. 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.

2. Summary

20,403 pet products were subcategorized in total. Each product had three labels; an animal category, big category, and 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 the result of the subcategorization, the products were classified into 74 kinds of groups (table 4). This table showed what the popular categories were. For example, 'treat' category in 'food, treat, treatment' big category of the dog products had almost same number of products as '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 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 4. 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