

The background of the slide features a textured, light gray wall. Three large, empty rectangular frames with thick black borders are arranged horizontally across the upper half of the image. The first two frames are partially obscured by a dark gray rectangular area that contains the title and subtitle. The third frame is on the right side of the slide.

PET PRODUCT AUTO-SUBCATEGORIZATION BY REVIEW ANALYSIS

MILESTONE REPORT

NAMIKO NAKASHIMA

TABLE OF CONTENT

1. Problem Statement

- 1.1. Problem
- 1.2. Client

2. Description of Dataset

- 2.1. Data Source
- 2.2. Data Cleaning
- 2.3. Data Wrangling

3. Findings from Exploratory Analysis

- 3.1. Tokens
- 3.2. Number of Products relating to each animal kind
- 3.3. Number of Products relating to each application type
- 3.4. Relationship between animal kinds and application types



1. PROBLEM STATEMENT

1.1. PROBLEM

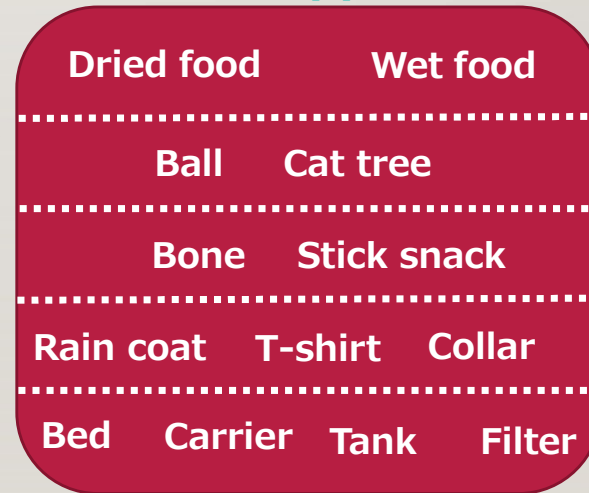
E-commerce companies set up **categories** for their products (e.g. Clothing, Beauty, Books, **Pet Supplies**, etc.)

If the number of products in a category has been growing...

Pet Supplies



Pet Supplies



Subcategory

← Food

← Toy

← Treat

← Clothes

← Other

Goal: Making a system **automatically subcategorizing** products by the **reviews**

1.2. CLIENT

E-commerce companies

By subcategorizing the products:

- Improving **analysis of trends** and **customer needs**
- Increasing **customer satisfaction** by easy access to a product they want

2. DESCRIPTION OF DATASET

2.1. DATA SOURCE

From [AWS](#)

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

	marketplace	product_id	product_title	product_category	review_body	review_date
0	US	510387886	(8-Pack) EZwhelp Belly Band/Wrap	Pet Products	Best belly bands on the market! These are a g...	2015-08-31
1	US	912374672	Warren Eckstein's Hugs & Kisses Vitamin Minera...	Pet Products	My dogs love Hugs and Kisses. However, the la...	2015-08-31
2	US	902215727	Tyson's True Chews Premium Jerky - 12 ounce Ch...	Pet Products	I have been purchasing these for a long time. ...	2015-08-31

Focused on the data collected during 2014-2015 in the US.

2.2. DATA CLEANING

1. Extracting necessary columns
2. Extracting data collected in **2014 or 2015**
3. Removing columns having a **single value**
4. Dropping with **missing values** and **duplicate data**



- 1,697,225 observations (reviews)
- 128,995 products

	product_id	product_title	review_body
0	510387886	(8-Pack) EZwhelp Belly Band/Wrap	Best belly bands on the market! These are a g...
1	912374672	Warren Eckstein's Hugs & Kisses Vitamin Minera...	My dogs love Hugs and Kisses. However, the la...
2	902215727	Tyson's True Chews Premium Jerky - 12 ounce Ch...	I have been purchasing these for a long time. ...
3	568880110	Soft Side Pet Crate, Navy/Tan	It is extremely well constructed, it is easy t...
4	692846826	EliteField 3-Door Folding Soft Dog Crate, Indo...	Worked really well. Very pleased with my purc...

2.3. DATA WRANGLING 1

1. Removing **short and long reviews**

Fewer than **30** characters ... less information
more than **760** characters ... too much information

2. Adjusting the **number of reviews** per product

Dataset 1: **2 to 5** reviews / product

Dataset 2: **5** reviews / product

Dataset 2: **5** reviews / product

How many reviews per product would be enough to subcategorize a product??

Not sure. Try **3 patterns!**

3. **Merging** reviews to the same product in **one** (For the next step, tokenization)

2.3. DATA WRANGLING 2

4. Tokenization

5. Removing

- **Non-alphabet** words
- **Stop words**

6. **Stemming** and lemmatization

7. Removing

- **Non-noun** words
- Words appearing **5** times or less

8. Removing products having **4 tokens** or less

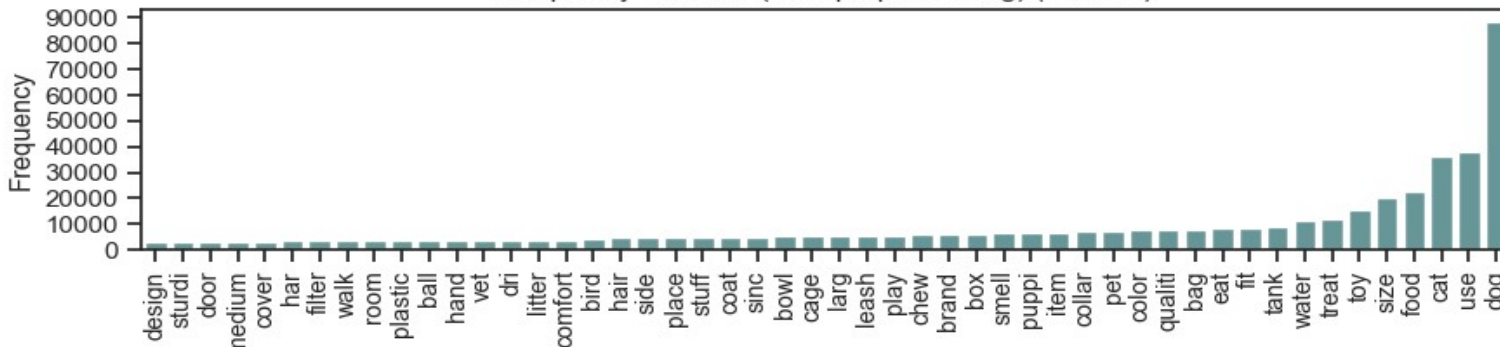


Dataset	Reviews / product	Total reviews	Total products	Total tokens (unique)
Original	various	1,697,225	128,977 (100%)	-
1	2 to 5	245,565	61,770 (48%)	8,032
2	5	161,765	32,295 (25%)	6,385
3	10	195,570	19,557 (15%)	6,840

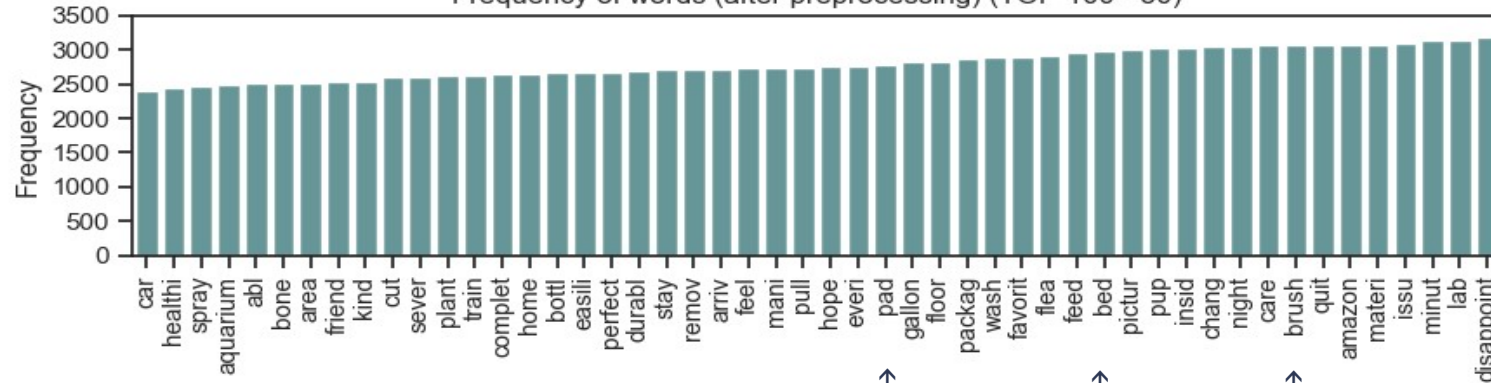
3. FINDINGS FROM EXPLORATORY ANALYSIS

3.1. TOKENS 1

Frequency of words (after preprocessing) (TOP 50)



Frequency of words (after preprocessing) (TOP 100 - 50)



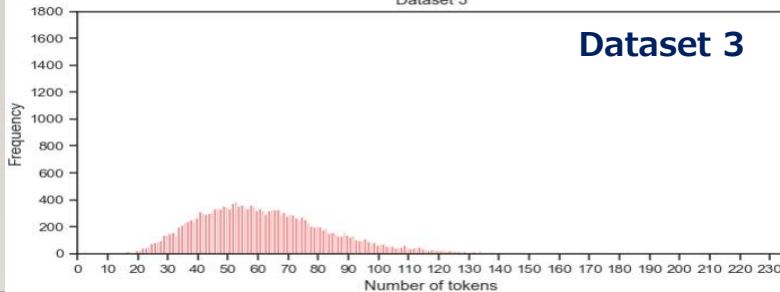
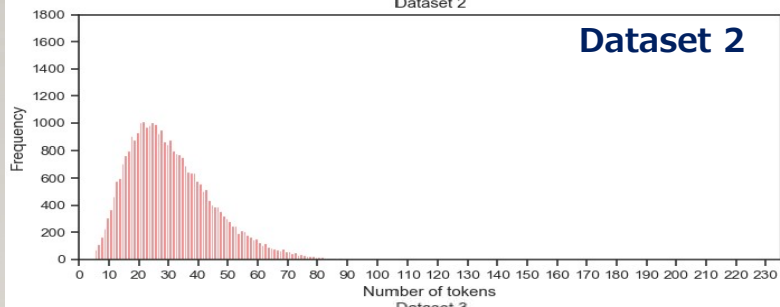
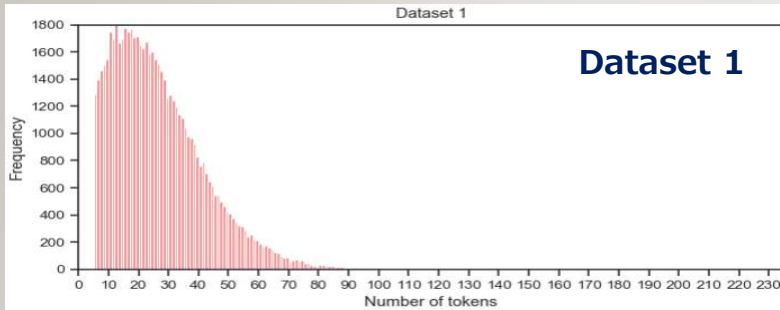
Words

* Dataset 1

↑ Animal kinds
↑ Application types

Will be a
significant clue to
subcategorize
products

3.1. TOKENS 2



Dataset	Min	25%	50%	75%	90%	Max	Mean	SD
1	5	14	23	35	47	119	26.2	15.8
2	5	20	28	39	50	130	30.5	14.6
3	9	46	59	76	94	234	62.5	24.0

The **more reviews** per product,
the **more tokens** per product

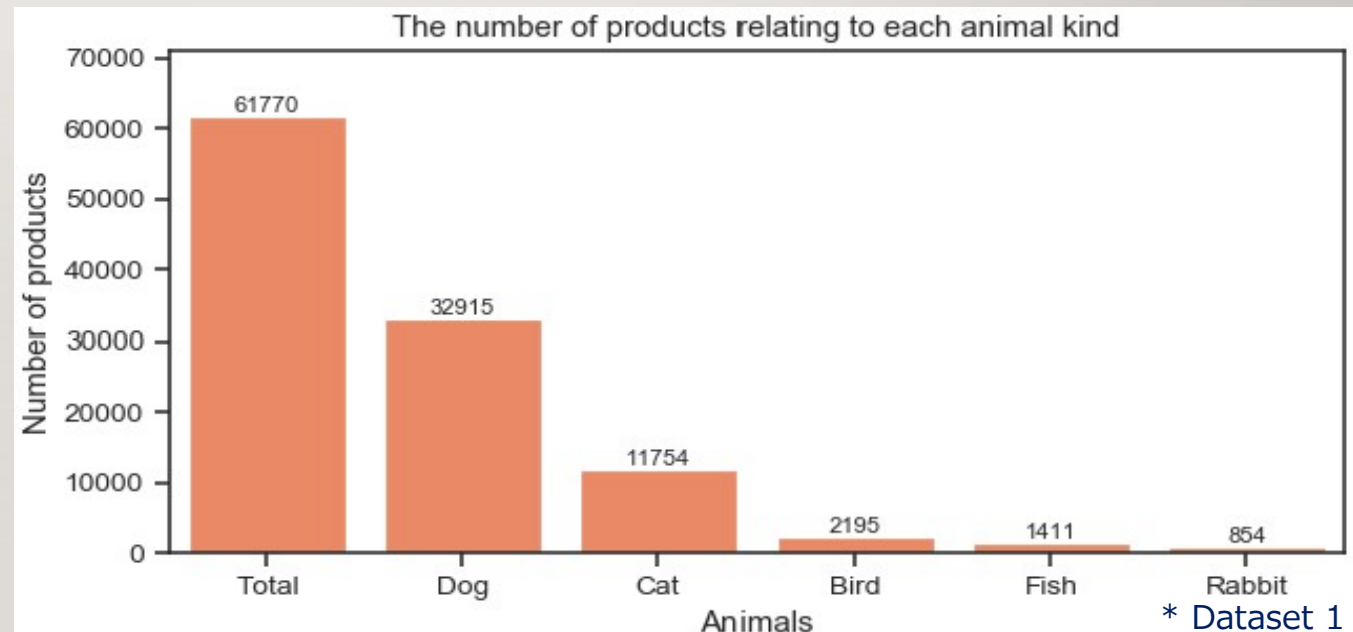
The number of tokens per product

3.2. NUMBER OF PRODUCTS RELATING TO EACH ANIMAL KIND

Categories: **Dog, Cat, Bird, Fish, Rabbit**
(e.g. Product having 'dog' as a token → Dog category)

The keywords for each category

Categories	Keywords
Dog	dog, puppi, doggi
Cat	cat, kitti, kitten
Bird	bird, chick
Fish	fish
Rabbit	rabbit, bunni



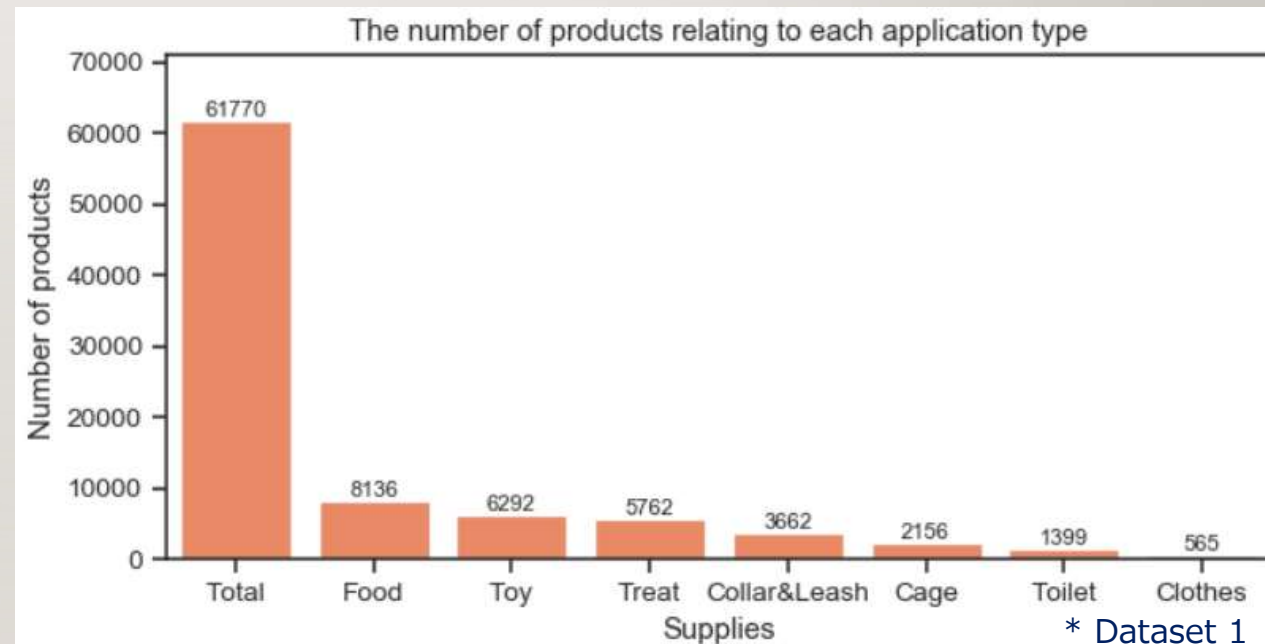
Dog and **Cat** were the largest two categories

3.3. NUMBER OF PRODUCTS RELATING TO EACH APPLICATION TYPE

Categories: **Toy, Food, Treat, Collar & Leash, Clothes, Cage, Toilet**
(e.g. Product having 'toy' as a token → Toy category)

The keywords for each category

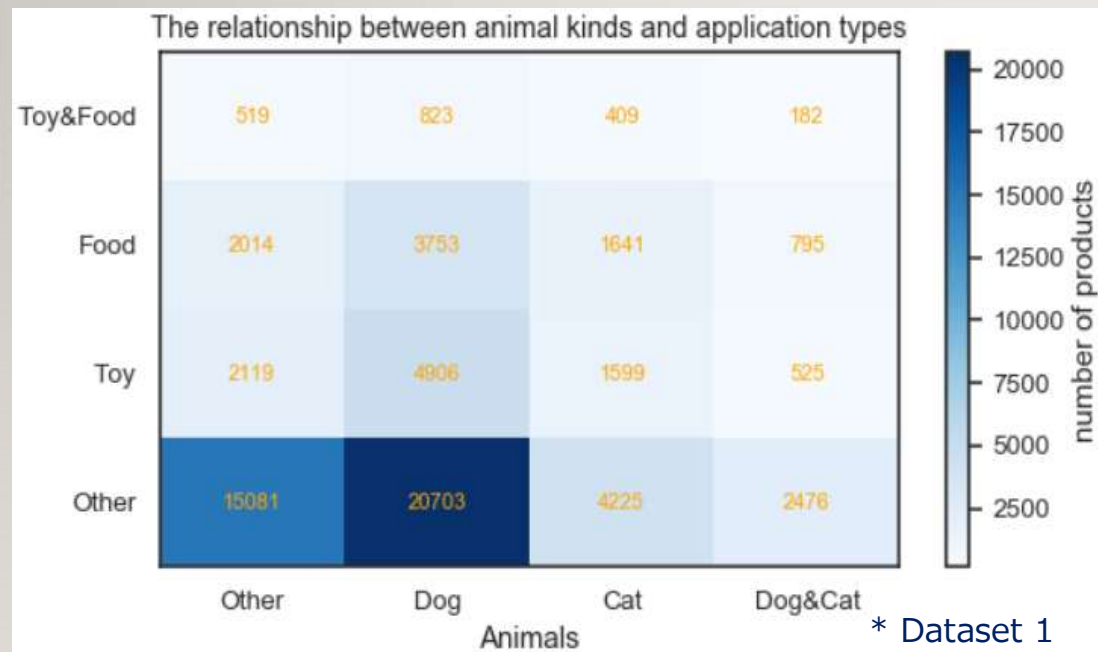
Categories	Keywords
Toy	toy, tunnel, ball, rope, stuff
Food	food, dri, wet
Treat	treat, snack, cooki
Collar & Leash	collar, leash
Clothes	shirt, coat, sweater, costum
Cage	cage, crate, carrier, kennel
Toilet	litter, pad



Toy and **Food** were the largest two categories, but relatively **spread**

3.4. RELATIONSHIP BETWEEN ANIMAL KINDS AND APPLICATION TYPES

Categories: [Dog, Cat] × [Toy, Food]
(e.g. Product having 'dog' and 'toy' as a token → Dog & Toy category)



- Food and Toy for dogs were the largest two categories
- Many products are in Other for dogs or Other for the other animals

Products will be categorized more precisely in the modeling part of this project.



THANK YOU

END