Capstone project 1: Statistical analysis report

The goal of this project is to create a system that automatically classifies products in Pet Supplies category into subcategories by analyzing the reviews. In this project, the data collected during 2014 - 2015 in the US is used.

1. The number of characters per review

There were 1,697,225 reviews after removing the missing values and duplicate data. The number of characters per review was analyzed as below.

Table 1. The number of characters per review

Min	25th percentile	50th percentile	75th percentile	95th percentile	Max	Mean
1	59	136	281	760	44,599	233.7

There were 477 reviews having only one character. The characters were a random alphabet/number or an emoji. 2,359 reviews had only two characters, e.g. 'ok', ':/' or 'A+'. The median was 136 characters. On the other hand, there were very long reviews having several-thousand or more characters.

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 would make the model complicated. Observations that had a short review (fewer than 30 characters) and long reviews (more than 760 characters, 95th percentile) were removed. In consequence, 12,177 (9.4%) products were dropped.

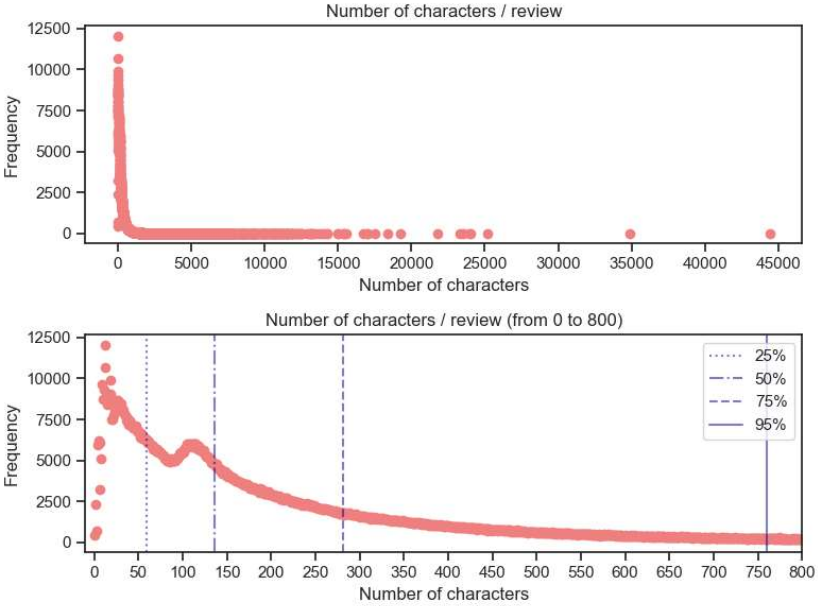


Figure 1. The number of characters per review

2. The number of reviews per product

There were 1,396,300 reviews and 116,800 products after selecting the reviews having 30 ≦ and ≦ 760 characters. Here, the number of reviews per product was analyzed as below.

Table 2. The number of reviews per product

Min	25%	50%	75%	95%	Max	Mean
1	1	2	5	42	4,285	12.0

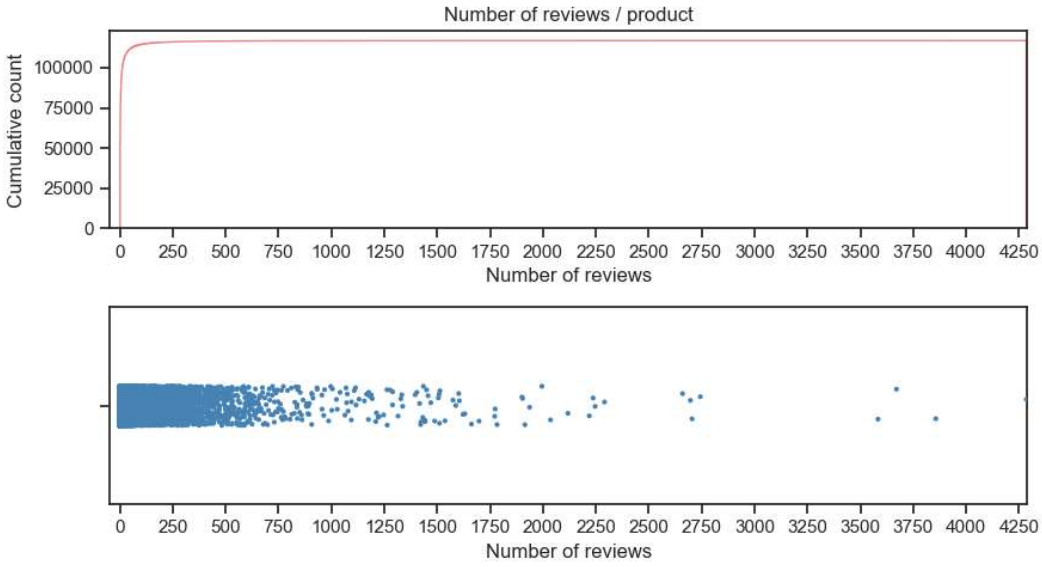


Figure 2. The number of reviews per product

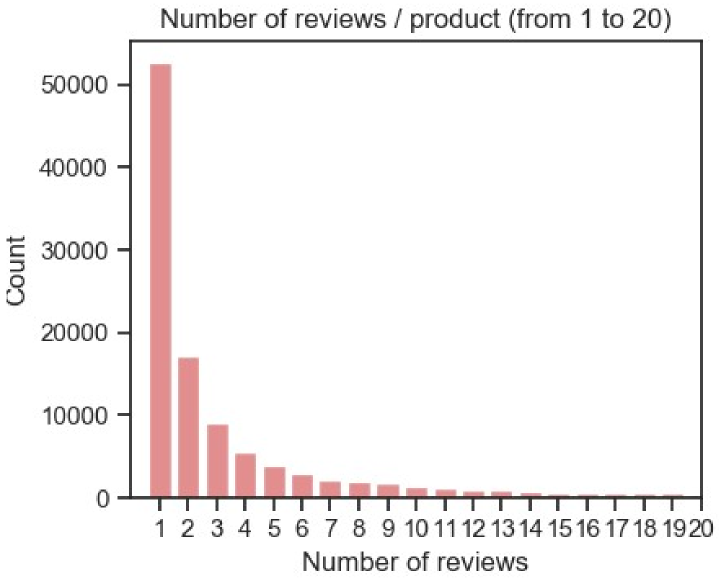


Figure 3. The number of reviews per product (from 1 to 20)

95% of products had less than 42 reviews although some had many. The range from 1 to 20 reviews was picked up as a count plot (Figure 3.) As Table 2 showed, the fewer number of reviews, the more frequent.

At this moment, I was not sure how many reviews per product would be enough to subcategorize a product. Of course, it depends on the length of a review and the number of characteristic words in a review. So, I decided to prepare three types of datasets depending on how many reviews a product had; 2 to 5 reviews per product (Dataset 1), 5 reviews (Dataset 2), and 10 reviews (Dataset 3).

3. The number of tokens per product

The reviews were divided into tokens, and the tokens were preprocessed to be used for the following analysis (see Data Wrangling Report for the detail.) The number of tokens per product was shown below for each Dataset.

Table 3. The summary statistics of the number of tokens per product

Dataset	Min	25%	50%	75%	90%	Max	Mean	SD
1	5	14	23	35	48	119	26.3	15.9
2	5	20	28	39	51	130	30.6	14.7
3	10	46	60	76	94	235	62.8	24.1

There were 61,796 (Dataset 1), 32,296 (Dataset 2), and 19,557 (Dataset 3) products in each Dataset after preprocessing. Because each Dataset had a different number of products, the heights of the bar plots were different (Figure 4). In Dataset 1, 90% of the products had 48 or fewer tokens, and the mean was 26.3. Datasets 2 and 3 had the larger number of both the 90th percentile and the mean. This was because Dataset 2 or Dataset 3 had 5 or 10 reviews per product, respectively, but 2 to 5 reviews in Dataset 1. Dataset 3 had a larger standard deviation.

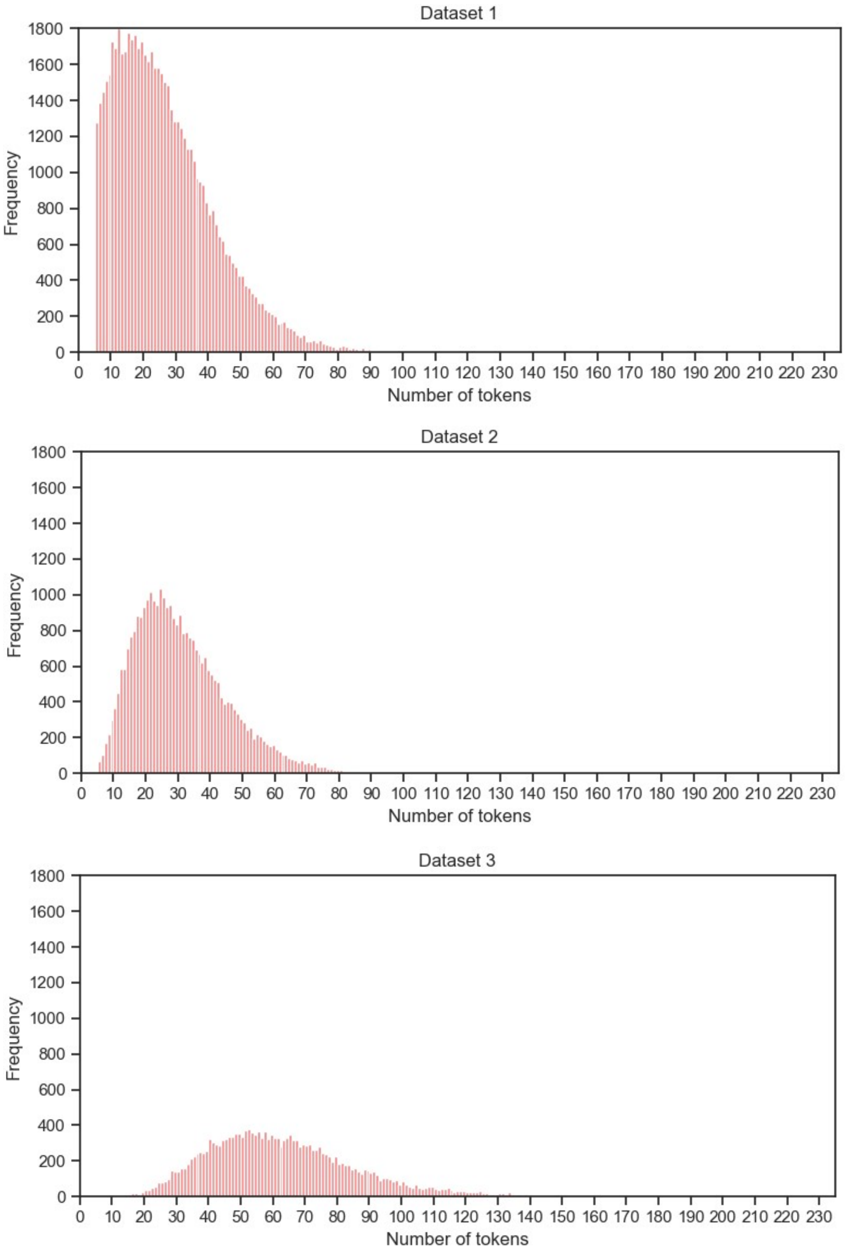
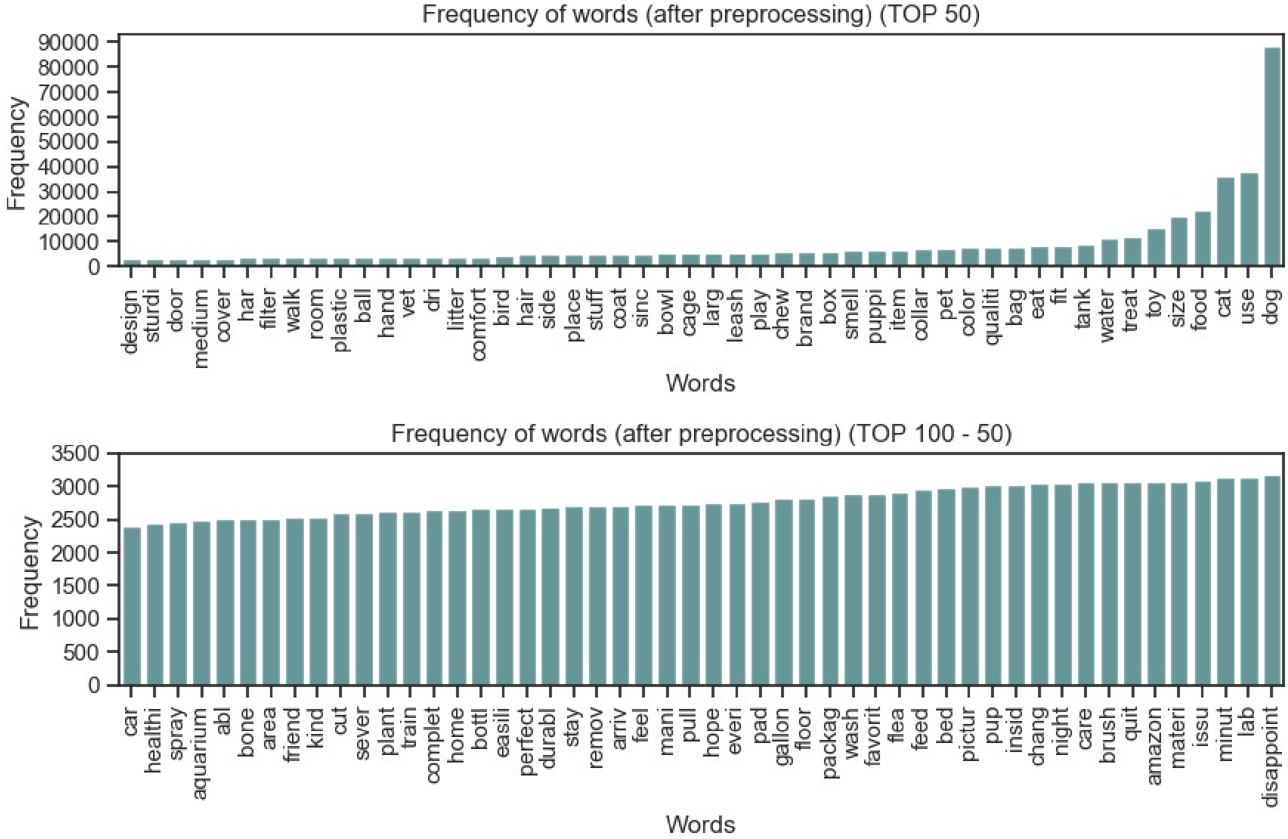


Figure 4. The number of tokens per product

4. The frequency of words in all reviews

There were 9,140 (Dataset 1), 7,278 (Dataset 2), and 7,872 (Dataset 3) unique words in each Dataset after preprocessing. Words (tokens) about animal kinds and application types appeared high up on the list. For example, animal kinds: dog, cat, and bird. Application kinds: food, toy, treat, tank, collar, cage, leash, bowl, etc. These words will be a significant clue to subcategorize products. I picked up Dataset 1 here (Figure 4), but the other two datasets were similar.


Frequency of words (after preprocessing) (TOP 50)

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

Figure 4. The frequency of words (Dataset 1)