Capstone project 1: Data wrangling 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. Cleaning and wrangling Steps:

The following steps of data cleaning and wrangling were performed:

- Extracting columns which were going to be used when loading the dataset,
- Extracting data collected in 2014 or 2015,
- Removing columns having a single value,
- Dealing with missing values and duplicate data,
- Removing short and long reviews,
- Adjusting the number of reviews per product,
- Merging reviews to the same product in one
- Tokenization,
- Retaining only alphabets,
- Removing stop words,
- Stemming and lemmatization,
- Retaining only nouns,
- Removing words appearing in 5 or fewer products, and
- Removing products having 4 tokens or less

2. How to deal with the missing values and duplicate data

There were 1,705,229 observations (reviews) and 128,995 products in the data at the beginning. 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 observations (reviews) remained, but the number of products in the data did not change (128, 977 products).

3. How to prepare for a data analysis

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. 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 makes the model complicated.

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). When a product had fewer reviews than the set amount, it was removed. When a product had more reviews than the set amount, the latest reviews were selected for the product. 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. First, non-alphabet words and stop words were removed. Then, the other 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 times 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 100 appearance that were not characteristic were removed by hand.

Also, products having fewer than 4 tokens after cleaning were removed. It was caused because some products had a few reviews and the reviews did not have enough characteristic information.

After this processing, the number of reviews and the products, and the number of unique tokens were following. Dataset 1 held 48% of the original number of products, but 25% (Dataset 2) and 15% (Dataset 3). When more products were expected to be subcategorized, Dataset 1 would be the best. On the other hand, Dataset 1 had fewer reviews per product, and I was not sure that the amount was enough. I will figure it out when they are applied to the model.

Table 1. Three types of datasets and the numbers

Dataset	Reviews / product	Total reviews	Total products	Total tokens (unique)
-	(Original*)	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

* After removing missing values and duplicate data

Just in case, all products that were removed from the data were shown as tables in the Jupyter Notebook.