Dog Rating Data Analysis – Temitayo Ilori

Dog rate consist of three datasets. twitter_archive_enhanced.csv consists of 5000+ dog rating data which was filtered to 2356 containing data on twitter id, source, url, name and the dog stage.

Tweet dataset scraped directly from Twitter consists of retweets that tracks retweet count, favorite counts among many other fields.

The third dataset contains image prediction (named predictions) consists of twitter id, top three predictions and their probabilities.

The data gathering steps used in this analysis included:

Gather

twitter_archive_enhanced.csv was provided for this analysis. I downloaded the file, uploaded it to the Udacity Jupyter project workspace. Then I loaded the dataset using read_csv and named it df. In order to preserve the original dataset while cleaning, I made a copy of this dataset and named it TwitterEA. I used this copy in the data analysis.

tweet_json.txt was scraped through Twitter API. I name this Tweet. In the same way I made it I say an I called it TwitterT, which I used for data analysis.

I downloaded image_predictions.tsv programmatically with the request module. Then, I loaded the dataset using read_csv and named it predictions. I also made a copy of this dataset and named it Twitter P.

Assess

I assessed the three datasets both visually and programmatically.

Visual assessment: My visually assessing TwitterEA, I discovered the extended_url column sometimes consists of multiple urls separated by commas. It also shows that the name column is often not correct, especially when the text field contains phrases like "this is a", "this is an", "this is such" and so on.

The visual assessment of TwitterT shows me that display_text_range is of the format [0, 85]. This needs to be broken down into display_text_minimum and display_text_maximum i.e 0 for display_text_minimum and 85 for display_text_maximum.

Programmatic Assessment:

- Showed me that part of the requirements is to only include favorite_count and retweet_count in this dataset. TwitterT.columns showed me all the columns in the dataset. I took note of this in order to determine which columns to drop.
- TwitterAE.info() showed me that Timestamp and Retweeted_status_timestamp are objects instead of datetime.
- TwitterAE['expanded_urls'] confirmed that some entries in the expanded_urls column have multiple values, separated by commas.

- TwitterAE['name'].value_counts() Some of the entries in the name column are incorrect. For example, "a", "the", "an" are cases where the text include "this is a", "this is the" and "this is an" respectively.
- TwitterAE['name'].sort_values() This shows more errors in the name column like "very", "unacceptable", "such", "quite", "officially", "not", "my" etc.
- Some name values start with capital letters while some don't
- TwitterAE.expanded_urls.value_counts() Some entries in the expanded_urls column are duplicates.
- TwitterAE[~TwitterAE['retweeted_status_id'].isnull()] there are a few tweets that are retweet (having @RT)
- The display_text_range in TwitterT should be broken down to display_text_minimum and display_text_maximum.
-

Cleaning Quality Issues

I took the following steps to clean the data:

1. Splited expanded_urls containing repeated urls, separated by commas. Leave only one url in the field.
2. Converted Timestamp and Retweeted_status_timestamp to datetime
3. Removed space before some expanded_urls
4. Replaced incorrect name entries with "None".
5. Caipitalized the first letter of each name.
6. Removed retweets
7. Removed duplicates expanded_urls columns
8. Removed duplicates from the prediction dataset.

Cleaning Tidiness Issues

1. Broke down display_text_range in TwitterT into display_text_minimum and display_text_maximum.
2. Merged Twitter archive (TwitterAE) and the prediction (TwitterP) datasets on tweet_id.