Wrangle and Analyze Data

By/ Mohammed Saleh
MohammedSaleh@ieee.org

Wrangle Report

14th December 2020

OVERVIEW

This report briefly describes your wrangling efforts during working on this project.

TOPICS

This report will cover:

1. Data Gathering.
2. Data assessing.
3. Data cleaning.

DATA GATHERING

There are three sources of data during this project:

1. The WeRateDogs Twitter archive: *"twitter_archive_enhanced.csv"* which was provided by Udacity and has been manually added.
2. The tweet image predictions, i.e., what breed of dog (or another object, animal, etc.) is present in each tweet according to a neural network. This file (*image_predictions.tsv*) is hosted on Udacity's servers and downloaded programmatically using the Requests library.
3. Each tweet's retweet count and favorite ("like") count at minimum, and other additional data. Using the tweet IDs in the WeRateDogs Twitter archive, and querying the Twitter API for each tweet's JSON data using Python's Tweepy library and storing each tweet's entire set of JSON data in a file called *tweet_json.txt* file.

DATA ASSESSING

Quality issues

1. *"twitter_archive_enhanced"* data:
    a. *in_reply_to_status_id*, *in_reply_to_user_id*, retweeted_status_timestamp, *retweeted_status_id* and *retweeted_status_user_id* are in float format and have a few inputs to analyze.
    b. *timestamp* is in a string format.
    c. The *313* row has a wrong *rating_numerator* and *rating_denominator*.
    d. There are decimal values for rating but the *rating_numerator* is an integer.
    e. The rows *1712*, *763*, *695* and *340* have a wrong *rating_numerator*.
    f. *source* column includes HTML tags.

2. *"tweet_json"* data:

    a. We don't need the retweets.

    b. After removing retweets the *retweeted_status* column will be useless.

Tidiness issues

1. Combine doggo, *floofer*, *pupper* and *puppo* into a single column called dog stage.
2. Merging three data sets into one.

DATA CLEANING

Programmatically cleaning

1. *“twitter_archive_enhanced”* data:
    a. Dropping all of the columns: *in_reply_to_status_id*, *in_reply_to_user_id*, retweeted_status_timestamp, *retweeted_status_id* and *retweeted_status_user_id*.
    b. Creating a new *year*, *month* and *day* columns from *timestamp* column.
    c. Removing HTML tags from *source* column.
    d. Converting *Doggo*, *Floofer*, *Pupper* and *Puppo* to boolean.
2. *“tweet_json”* data:
    a. Dropping the retweets and *retweeted_status* column.
    b. Converting *tweet_id* column into the integer type.

Manually cleaning

1. *“twitter_archive_enhanced”* data:
    a. Changing the values of *rating_numerator* and *rating_denominator* in row 313.
    b. Changing the values of *rating_numerator* for the rows *1712*, *763*, *695* and *340*.

And finally, merging the whole data into one data file and dropping rows with NULL values.