Wrangle Report

By ALI IRTAZA

Introduction

The dataset that you will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

WeRateDogs downloaded their Twitter archive and sent it to Udacity via email exclusively for you to use in this project. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017. More on this soon.

Gathering Data

We will gather the data from three resources:

1. The twitter_archive_enhanced.csv file is provided to me.
2. The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. The file is provided.
3. I will use the tweet-json.txt file to gather each tweet's retweet count and favorite ("like") count at minimum.

Assessing Data

We have to detect and document at least eight (8) quality issues and two (2) tidiness issues

Issues Found

Quality

1. Archive

There are missing data in the following columns:

1.in_reply_to_status_id

2.in_reply_to_user_id

3.retweeted_status_id

4.retweeted_status_user_id

5.retweeted_status_timestamp

- There are extra columns that aren't useful for the analysis(Some are above columns stated). They should be deleted from the data set.
- The numerator and denominator should be float not integer
- The value of denominator should be 10 only but there are multiples of values too.
- There are some incorrect numerator rating values eg 1776, 960, 666 etc.
- A standard column of Rating should be there that shows the truely value of the dogs rating
- The column timestamp is object type where it should be timestamp
- There are some invalid names of dogs that should should be corrected.

2.Image Prediction

- There are 66 duplicated urls that should be removed from the data
- There are invalid data in column p1,p2,p3 like ibex ,bagel, fruits name like banana and etc
- There are extra columns that aren't useful for the analysis
- There should be one column for prediction and one column for the confidence

3. Twitter API Data

- There are some missing values in the data

Tidiness

- The last four columns in archive data ie doggo, floofer, pupper, puppo should not have seperate columns,they should have one column as every tweet_id should have any one observation from these type.
- All tables should be part of one data set

Cleaning Data:

Define

1. The four different types of dog should be in a same column
2. The column timestamp is object type where it should be timestamp
3. Convert Numerator Rating and Denominator rating from int to float in archive
4. Correcting naming issues of the dog
5. Deleting duplicates jpg_url from image prediction data
6. Creating 1 column for Prediction and 1 column for Confidence in image prediction data
7. Deleting extra columns from the image prediction data set
8. Merge all three data sets
9. Deleting retweets from the data
10. Removing all the columns that are not useful for analysis
11. Convert tweetid from integer to string
12. Creating a standard 'rating' column from numerator and denominator ratings that shows the true value of rating