WeRateDogs Twitter Archive – Wrangle Report

By Seif El-Eslam Hegazy

This report outlines the wrangling efforts to assemble and clean the data required for analysis of the WeRateDogs Twitter Archive.

Data Gathering

Data is gathered from 3 sources, stored in separate files:

1. WeRateDogs Twitter Enhanced archive, manually downloaded from the Udacity servers.

2. The image predictions file, programmatically downloaded from the Udacity servers.

3. The entire set of each tweets' JSON data, downloaded by querying the Twitter API using the Tweepy library. The favourite_count and retweet_count were extracted programmatically from this file.

These 3 raw data files are loaded into separate data-frames: **df_twitter_archive, df_image_prediction and df_tweet_data**.

Assessment & Cleaning

Assessment begins by viewing the information on the **df_twitter_archive** data-frame first, identifying several quality and
tidiness issues.

All rows containing non-null values in the **retweeted_status_id** , **retweeted_status_user_id** and **retweeted_status_timestamp**, and also in the **in_reply_to_status_id** and **in_reply_to_user_id** columns are dropped, as per the requirements. These columns are then also dropped.

The **timestamp column** is converted to datetime data type. Month and day of week are extracted for further analysis.

The 4 dog stage columns are melted into the **stage column**; tweets without stages are set to ‘none’.

The html strings in the **source column** are replaced with the display portion of itself.

The **rating_numerator** and **rating_denominator** columns are checked for value ranges; It is decided to keep only tweets with single ratings.

Tweets with large numerators are dropped, as the text didn’t contain a valid rating (# out of 10). After the ratings are fixed, rating_denominator column (it contained only ‘10’s) is dropped and **rating_numerator column is renamed to rating**.

The odd words in the **name column** are replaced with ‘none’.

Tweets with missing values in **expanded_urls** , (not retweets or replies) are actually missing the urls from the text itself. These tweets are dropped.

The predictions dataframe itself is not cleaned. There are many tweets with no dog breed predicted, these rows are dropped. The best prediction for **breed and associated confidence level are extracted and merged into the df_twitter_archive data-frame**.

The json_data table itself was not cleaned. **The retweet_count and favourite_count columns were merged into the df_twitter_archive**.

The remaining cleaned data in the df_twitter_archive dataframe are saved to the new “**twitter_archive_master.csv**” file. The predictions and json_data tables had not been cleaned, so were not saved.