Data Wrangling

Background

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. It was started in 2015 by college student Matt Nelson, and has received international media coverage both for its popularity and for the attention drawn to social media copyright law when it was suspended by Twitter.

Gathering data

We first gather three pieces of data

1. **Twitter_achieve_enhanced.csv**, we can download this file from WeRateDog Twitter archive.
2. The tweet image predictions, i.e., what breed of dog is present in each tweet according to a neural network. This file (**image_predictions.tsv**) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv
3. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called **tweet_json.txt** file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count.

Assessing

After gathering the data, we can analysis the data quality and tidiness.

Quality

In **Twitter_achieve_enhanced.csv**, ie: arch_df (Twitter archive) table, we found

- Incorrect datatype of **tweet_ID**.
- Incorrect datatype of **in_reply_to_status_id,in_reply_to_user_id**
- contains retweets, as we only want original ratings (no retweets). So the retweets shouldn't be there
- duplicates many **tweet_id(s)** of arch_df table, in which they are missing in image_df (image predictions) table. We only want ratings with images. But not all ratings have images.
- incorrect datatypes (**in_reply_to_status_id**, **in_reply_to_user_id** and timestamp columns) for timestamp
- unnecessary html tags in source column in place of utility name e.g. <a href=""http://twitter.com/download/iphone"" rel=""nofollow"">Twitter for iPhone
- text column contains untruncated text instead of displayable text
- rating_numerator column has values less than 10 as well as some very large numbers (e.g. 1176)
- rating_denominator column has values other than 10
- incorrect dog names starting with lowercase characters (e.g. a, an, actually, by)
- some records have more than one dog stage
- nulls represented as 'None' in columns 'name', 'doggo', 'floofer', 'pupper','puppo'.
- some dog names are incorrect.

Tidiness

- **arch_df** remove any duplicates **retweeted_status_id, retweeted_status_user_id** and **retweeted_status_timestamp** columns.
- doggo, floofer, pupper and puppo columns in arch_df table could be merged into one column named as **stage**.
- **breed** column could be added in arc_df table as its values based on p1_conf and p1_dog columns of image_df (image predictions) table.
- retweet_count and favorite_count columns from status_df (tweet status) table should be joined with arch_df table

Cleaning

- Based on above Quality and Tidiness, we start to clean the dataset.
- To achieve our goals and meet expectations, we must first determine how will we execute data cleanup successfully.
- Much of data cleaning can be done by manual and/or software. However, manual process is an overwhelming task. Whereas software cleaning process must be monitored carefully. Therefore, it is better to build a protocol for data cleaning.

Storing and Analyze

After our data has been standardized, validated, and cleaned. We save this master data as **twitter_arch_master.csv**

Then, we can start to analyze the data.

Conclusions

We knew that Data quality is important. Old and inaccurate data can have an impact on analyzing results. While Data cleaning, is the process of ensuring that our data is correct, consistent and useable by identifying any errors or corruptions in the data, correcting or deleting them, or manually processing them as needed. But using tools to cleanup data will make more efficient.