Reporting: A Wragle Report on WeRateDogs Tweet Archive

Gathering Data

The data was gathered from 3 different sources:

1. A file “twitter-archive-enhanced.csv” downloaded from the link in the classroom.
2. A file “image-predictions.tsv” downloaded using the requests library
3. Data gotten from Twitter API (with tweet id from the twitter-archive-enhanced.csv file) using Tweepy library.

Below is a report of the issues I found in the data and how I fixed them.

Quality issues

1. 181 tweets that are retweets and we don't need them.

    We fixed this by fishing out those tweets and dropping them. The tweets can be identified by checking for tweets that have retweeted_status information e.g retweeted_status_id

2. The source column contains the device the tweet was sent from but it is contained in a HTML element.

    We fixed this using regular expressions to match the closing and opening tags of the HTML elements in this column and removing them, thereafter using the split function to separate the “Twitter for” from the device name “iPhone”, for instance.

3. Some columns are not necessary.

    We fixed this by dropping the columns.

4. The timestamp contains date and time. We need to separate them in case we need to compare tweets based on the times and days of the week.

    We first converted the timestamp from object type to datetime type using Pandas’ to_datetime method. Then we picked the date part and the time part separately.

5. Wrong ratings.

    We worked with the text column to get the right ratings. First we used regular expressions to pick out the fraction representing the ratings. Then we split them by the / symbol and converted them to integers. Then we used these values to replace the ratings numerator and denominator.

6. Rating numerator higher than 20 with a denominator in multiples of 10.

We fixed this while fixing the wrong ratings problem. We reduced the ratings to a fraction with a denominator of 10.

7. rating_numerator with a high value of 1776 and rating_denominator value of 10

    We checked the text of these tweets and saw that they weren’t dogs, so we dropped them.

8. Some of the tweets are not for dogs.

    We first dropped all the tweets that were false for all 3 predictions. Then, for the remaining tweets, we picked the prediction that is both true and of the highest confidence (if there are other true predictions).

Tidiness issues

1. The tweets data we need are in 3 tables.

    We merged these 3 tables based on the tweet_ids, using the Pandas’ merge method.

2. Dog stages are in 3 different columns

    We fixed this using the Pandas’ melt function to move the dog stage data into one column.