wrangle_report

January 10, 2018

1 Gather

Data were collected from three different sources.

First data was collected from the "twitter-archive-enhanced.csv" file which was in the same directory in which project notebook was located. The csv file was imported into pandas dataframe. The dataframe was named "archive".

Second data was extracted programmatically from a URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv. Python's request library was used to extract data from URL. The URL was split using "/" as the separator and the last value was the file name. This file was written in the content of our request. Then this file was imported as a dataframe in pandas using tab as the separator. The dataframe was named "images".

The third data was extracted from Twitter API using python's tweepy library. I needed to extract the favourites and retweet counts for each tweet. This data was then saved as a JSON file using UTF-8 encoding.

The images dataframe, the JSON file and the archive data were merged into a single dataframe. A copy of this merged data was saved in CSV format.

2 Assess

Pandas' `.info()` method on the merged data showed that timestamp column needed to be a datetime object instead of a string.

There were several empty values in *in_reply_to_status*, *in_reply_to_user_id*, *retweeted_status_id*, *retweeted_status_user_id*, *retweeted_status_timestamp*.

The *name* column had several entries which do not look like a name.

Some of the ratings did not look right. The expected value for numerator and denominator was around 10, but there were many values above 100 also.

The number of rows in archive data and images data did not match.

In several columns, the null values are treated as non-null values. Some entries contain "Nan" as string.

The *Unnamed: 0* column was to be removed.

The columns for dog breed predictions could be condensed into a single column.

The dog stages values were named as columns instead of one column containing stages values.

3 Clean

The extra column *Unnamed: 0* was removed.

Timestamp column was converted to datetime object.

Retweets were removed and tweets which did not include images were also removed because those tweets were not dog ratings.

Dog_type column was created which showed the type of dog(dog stages).

The dog breed predictions were converted into a single column.

After doing all these cleaning processes some useless columns were removed.

Ratings for dogs were extracted from the *text* column.

The *text* column was used to create names column.

After cleaning, the data was exported to a CSV file named "twitter_archive_master.csv"