WeRateDog Twitter Dataset – Wrangle Report

Created by Justice Annan

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

Data Gathering

The data used for this project consisted of three (3) different datasets, stored in separate files.

Twitter Archive Enhanced file: This dataset was provided by Udacity in the project guideline. It was manually downloaded from the Udacity servers and imported into a jupyter workspace by using the pandas library as pd and the read_csv() function to read the file into the workspace. It was then named as **df_enhanced.**

The image predictions file: this file was hosted on Udacity servers. It was programmatically downloaded using the Request Library from the following url: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

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

The three (3) raw data files were then loaded into separate tables: df_enhanced, df_image and tweet_json.

Assessment

Once the 3 datasets were obtained, they were assessed as follows:

Visual Assessment: all the 3 datasets were printed into the jupyter notebook using .head() and .tail() functions so they could be assessed visually. On the jupyter notebook, I scrolled left and right, up and down to visually assess the 3 datasets for quality and tidiness issues. The datasets were also opened with Microsoft excel to visually assess them for issues.

Programmatic Assessment: For programmatic assessment, I used methods and functions such as .info(), .value_counts(), .value_counts().head(), .str.islower(), len(), and type().

Cleaning Data

This section was divided into 3: Define, Code and Test.

Before the actual cleaning was started, a copy of the original 3 datasets were made into df_enhanced = enhanced_clean, df_image = image_clean and tweet_json = tweet_json_clean.

They were then followed by the processes for Define, Code, and Test to clean the copied datasets as follows:

All invalid names (lowercase names) were replaced with the string ‘None’. To be able to do this, I had to get all the names with lower case and afterwards, they were replaced with None.

All rows in the in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id**,** and retweeted_status_timestamp were dropped as they were not relevant in the analysis. Eventually the above columns were dropped since they had no data.

Timestamp datatype was converted from string to datetime datatype.

Tweets with missing data in the expanded_urls were dropped.

All four columns for doggo, floofer, pupper and puppo were combined into the stage column.

The json_data table as well as the image predictions table were all merged with the twitter archive enhanced table on tweet_id column.

Tweets with missing data for retweet, favorite and images were also dropped.

Storing Data

After the wrangling process (gathering, assessing, and cleaning data), the cleaned and merged data were saved to a csv file named twitter_archive_master.csv

Conclusion

The above shows the steps which were taken to successfully gather, assess and clean the 3 different datasets.