WeRateDogs Tweets Archive

By Muhammed Jimoh

18th February 2023

The method used to wrangle the WeRateDogs Twitter archive is summarized in this report. The steps involved: Gathering, Assessing, and Cleaning Data.

Gathering Data

The initial step in the process of data analysis is data gathering. Only occasionally is the data accessible in a single location. It typically needs to be acquired from various sources because it is dispersed. The information for this project was acquired from three separate sources.

- The WeRateDogs Twitter archive data was downloaded from Udacity.
- The image dataset, which was downloaded remotely from Udacity servers
- The Tweepy library was used to extract tweet data in JSON format from the Twitter API.

Assessing Data

The process of evaluating the data to identify organization and quality problems in the existing data is known as data assessment. The results of the individual assessments made for the datasets are listed below:

Quality

- timestamp, retweeted_status_timestamp column type should be date instead of object
- in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id data types should be int instead of float
- tweet_id: The current type is int64, I will convert it to string since there is no calculation required.
- created_at: The current type is object (str) rather than datetime. I will change it to datetime.
- source column includes HTML tags, these tags should be stripped.
- names column should be cleaned, there are invalid records like a, the, an, the, very, and unacceptable which start with lowercase.
- p1, p2, and p3 columns should be standardized "-" expression should be removed.
- for the archive data, retweeted data indicate duplicated tweets.
- some value of rating_numerator is over the scale and rating_denominator is greater than 10, normalize this.
- drop the following columns in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_timestamp, retweeted_status_user_id, rating_numerator, and rating_denominator.

Tidiness

- - joining all tables.
- - creating final dog prediction (with doggo, floofer, pupper, puppo in one column) column.

Cleaning Data

While cleaning the dataset, all of the problems mentioned above for the various datasets were resolved.

Though the aforementioned examination and cleaning technique may not have covered all the issues, we can take data cleaning a step further.