WERATEDOGS TWITTER ARCHIVE - WRANGLE REPORT

Akshai Ramesh
August 2019

This project mainly focuses on data wrangling to fix the data quality and tidiness issues using python.

DATA GATHERING

1. 'Twitter_archive_df' : The WeRateDogs Twitter archive, which is provided by Udacity and pd.read_csv() to import them into dataframe.
2. 'Image_df' : The tweet image predictions, i.e., what breed of dog (or other objects, animal, etc) is present in each tweet according to a neural network. This file('image_predictions.tsv') is hosted on Udacity's servers and downloaded programmatically using the requests library and the provided url.
3. 'Twitter_json_df' : 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 twitter_json_df file. Each tweet's JSON data is written to its own line.

DATA ASSESSMENT

The dataset is assessed for two types of issue : tidiness and quality.

Tidiness issues:

1)Merge the dog stages to a single column
2) Master dataframe should be created from the data present across three dataframes.
3) Numerator and denominator are present as separate column.

Quality issues:

1) Re-tweeted records should be removed

2) 'in_reply_to_status_id','in_reply_to_user_id' columns should be dropped.

3) Tweets posted after Aug 1 , 2017 should be removed.
4)Change the datatype of timestamp field from string to datetime.
5)Tweets with no images should be dropped.
6)Convert tweet_id from int to string.
7)Invalid dog names should be replaced with None.
8)Clean the denominator that has strange values.
9) Correct the 'rating_numerator' values from the text information.
10) Optimize the source content by ‘Twitter for iPhone’, ‘Twitter Web Client’, and ‘TweetDeck’.
11) Change datatypes of columns to their appropriate ones.