Mehmet Saruhan
08/12/2018

Data Wrangle Report for Tweet Data Project

This project aims to gather, assess, clean the data and produce insights as well as visualization of the data. Results are expected as a clean dataset composed of twitter data of user Weratedogs.

Gathered Files

1- image_predictions.tsv – Project website
2- twitter_archive_enhanced.csv – Project website
3- Tweet_json.txt – Gathered form Twitter API with tweepy

After assessing the data, I figured out some quality issues about data as

twitter_archive_enhanced

- Some numerators are missing
- We need a "like" value for analysis instead of numerator and denominator
- Source column should be cleaned and converted to category
- Dog names have errors a, officially ...
- There are missing dog names
- There are retweets
- Breed types should be a category
- tweet_id , in_reply_to_status_id,in_reply_to_user_id columns should be in string format
- timestamp and retweeted_status_timestamp columns should be in date format
- Merged image prediction column names are not understandable
- Column names start with lower letters
- Dog Stages have duplicated values
- Dog Stages have missing values
- Dog Stage values contain "None" which should be NULL

image-preddictions

- Some predictions in the same pictures have close probability. If 1 p_dog == false and 2nd and 3rd prediction probability is close and their p_dog == true they can be considered more precise.

Already extracted dog names and stages contain many errors. I figured out that all names start with capital letters. After assessing errors, name phrases clarified as

- "This is NAME"
- "named NAME"
- "hello to NAME"
- "Meet NAME"
- "name is NAME"
-

In addition to names, there were some missing and multiple dog stages. Stage reextracted from tweet text and tweets contains multiple stages excluded

I decided to use points rather than denominators and nominators for better analyze. Fixed wrong extracted numbers and get the value of nominator/denominator and stored as Point.

Finally, all 3 data files combined with a single python data frame. Retweet and tweet count from tweet_json.txt and Breed, isDog and Probability information from image_predictions.tsv data are combined into the main data frame.

https://stackoverflow.com/questions/23307301/pandas-replacing-column-values-in-dataframe
https://chrisalbon.com/python/data_wrangling/pandas_delete_duplicates/
https://stackoverflow.com/questions/34830597/pandas-melt-function