Wrangle Report

WeRateDogs

by Apurva Verma

Introduction

WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. The account was started in 2015 by college student Matt Nelson, and has received international media attention both for its popularity and for the attention drawn to social media copyright law when it was suspended by Twitter for breaking these aforementioned laws. The account's language has spawned an Internet language about "doggos" and "puppers". A 2016 interaction with another Twitter user, when Nelson purposefully misnamed him "Brent" as is common in Weird Twitter, spawned the catchphrase "They're good dogs, Brent", which became one of the biggest memes of 2016.

We have wrangle the twitter data through the following process.

1. Gathering data
2. Assessing data
3. Cleaning data

Then visualized the wrangled data.

Gathering

Twitter archive file: download this file manually by clicking the following link: twitter_archive_enhanced.csv

The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Twitter API & JSON: Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. 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 tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count.

Assessing data

We have assessed our gathered data and here are some quality and tidiness issues in the dataframe**.**

Quality Issues

For Twitter archived dataframe:

1. Tweet_id is an integer, it should be str.
2. timestamp should be in datetime not object.
3. drop columns that are not useful as 'retweeted_status_id', 'retweeted_status_user_id','retweeted_status_timestamp','in_reply_to_status_id', 'in_reply_to_user_id'.
4. To categorized sources in readable format
5. Name column have invalid names i.e 'such', 'a', 'quite', 'not', 'one', 'incredibly', 'mad', 'an', 'very', 'just', 'my', 'his', 'actually', 'getting', 'this', 'unacceptable', 'all', 'old', 'infuriating', 'the', 'by', 'officially', 'life', 'light', 'space'.

For tweet image predictions :

1. To correct the p1,p2 and p3 columns for consistency
2. To delete duplicated image url

3. tweet_id is an integer. should be str
4. There are 2356 tweets in the twitter1 dataframe and 2075 rows in the images dataframe. This could mean that there is missing data, or that not all 2356 of the tweets had pictures.

for Twitter API & JSON dataframe:

1. tweet_id is an integer
2. There are 24 missing tweets compared to the twitter_archive dataframe

Tidiness Issues

twitter_archive dataframe

1. one column for 'doggo', 'floofer', 'pupper', 'puppo'.
2. to create a column for dogs rating.
3. to create a column for dogs gender.
   images dataframe
4. to create a new column 'breed' with p1_dog,p2_dog and p3_dog.The columns for dog breed predictions can be condensed.
5. To create a new column for confidence interval
   twitter dataframe
6. 24 twitter_id informations are missing

Combine twitter_archive dataframe, images dataframe and twitter dataframe to one dataframe.

Cleaning

First we created a copy of each dataframes for cleaning purpose.This part of data wrangling was divided into three parts as on each mentioned issue:

1. Define
2. Code
3. Test

All the quality and tidiness issues has been solved here and a new dataframe named as twitter_archive_master.csv is made Combining twitter_archive dataframe, images dataframe and twitter dataframe.