wrangle_report

January 11, 2019

1 Gather

Data were collected from three following different resources.

1) "twitter-archive-enhaved.csv" - This csv file contained information about tweet id, dog name, dog stage, rating and the source. This file was manually imported into the pandas dataframe, which was named as "twitter_archive"

2) "image-predictions.tsv" - This file was programmatically downloaded from the website, 'https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv'. Python's request libaray was used to obtain the website content. Later, the file was stored into pandas dataframe as "image_predictions". This dataframe contained tweet_id, image url, image counts and dog_breed using the neural netwrok.

3) Twitter API - Tweepy library was used to query Twitter's API. This was done to obtain data about retweet count and favorite count, corresponding to various tweet ids. JSON data was transferred to "tweet_json.txt" file with each tweet's JSON data on its own line. This file was saved into "tweet_info" dataframe.

2 Assess

Pandas' various methods such as .info(), .value_counts(), .duplicated() were used to assess three different dataset.

1) twitter_archive_master:

- There were 2356 datapoints in this dataset.
- Following columns had many missing values: in_reply_to_status_id and in_reply_to_user_id, retweet_status_id, retweeted_status_user_id, retweeted_status_timestamp and expanded urls
- Datatype for timestamp was not in datetime format.
- The numerator and denominator rating was in in integer, instead of float. Also, these ratings contain many outliers.
- Four different types of dog stages were divided into four different columns, instead of having it under one column.
- Some of the dog names did not seem correct.

2) image_predictions:

- There were 2075 datapoints in this dataset.
- There were three differnt predictions were made using the algorithm to predict the dog breed from the image. The first prediction of the algorithm seem to have the most possible image.
- This dataset contained 324 entries, which did not represent a tweet about a dog.

3) Three dataframes should be joined into one master dataset, to have various information about each tweet id under one roof.

3 Clean

Following cleaning steps were implemented. Before cleaning, the copies of all three datasets were made to avoid tempering with the original datasets.

3.0.1 Quality:

- twitter_archive_clean:

  1) Removed rows with "retweeted_status_x' since we are only interested in original tweets only.
  2) Dropped the following columns from twitter_archive: "in_reply_to_status_id", "in_reply_to_user_id", "retweeted_status_id", "retweeted_status_user_id", "retweeted_status_timestamp".
  3) Converted the twitter_archive timestamp to datetime. Then parsed the column into date and time.
  4) Converted url to text from Source column of the twitter_archive dataframe.
  5) Droped the rating_denominator and changed the rating_numerator column to rating. Identified the outliers with rating and eliminate them.
  6) Changed non-dog names to null in the twitter_archive_clean dataframe. Also, capitalize the first letter of the name.

- image_predictions_clean:

  1) Droped rows, which contains p1_dog,p2_dog, and p3_dog as false
  2) Converted p1,p1_conf,p1_dog to more descriptive column names
  3) Droped other unnecessary columnns such as p2', 'p2_conf', 'p2_dog', 'p3', 'p3_conf', 'p3_dog
  4) Capitalized the breed name in p1
  5) Changed breed to categorical type

3.0.2 Tidiness

1) Created a column, "dog stage" by extracting values from these four columns (doggo, floofer, pupper, puppo) in the twitter_archive_clean dataframe. In the end, eliminate these four columns (doggo, floofer, pupper, puppo).

2) Join the three dataframes (twitter_archive, image_predictions, and tweet_info) into one master dataset via inner join based on tweet_id. This master dataset was named "twitter_archive_master".

4 Store

twitter_archive_master dataframe was stored into twitter_archive_master.csv file using pandas.