wrangle_report

June 27, 2020

1 Project: Data Wrangling with Twitter data

In this project I was working with the data from the Twitter account WeRateDogs that rates other people's dogs in a funny way. Later I combined tweets data with image prediction data made on photos from the posts.

First, I gathered data from 3 different places: - tweets archive data was downloaded from csv file - tweet image predictions data was downloaded as tsv file from Udacity's servers programmatically using the Requests library - additional tweet data (retweet count and favorite count) was downloaded from Twitter using Tweepy to query Twitter's API. Using tweet_id from archive data I requested tweet data from Twitter programmatically in Json format and stored the date in tweet_json.txt file. Later I downloaded data back and gathered retweet count and favorite count into the panda dataframe.

Second, I assessed gathered data both visually and programmatically for all three datasets separately.

Then, I cleaned the following issues:

1.0.1 Quality:

df_archive_copy table

1) Columns to delete: *timestamp, source, expanded_urls, name* - this data is not needed for analysis
2) Delete replies and retweets Some tweets in the archive dataset are retweets or replies, they do not have dog rating data. Delete records with not null data in these columns:

- *in_reply_to_status_id*
- *in_reply_to_user_id*
- *retweeted_status_id*
- *retweeted_status_user_id*
- *retweeted_status_timestamp*

Drop those columns, because they are all equal null for original tweets records

3) Clean *rating_denominator* column

- Set *rating_denominator* = 10 where rating_denominator !=10 but text column has correct rating
- Records with *rating_denominator* !=10 are not ratings of one dog, drop these records
- Drop *rating_denominator* column, because it has 10 in all rows

4) Clean *rating_numerator* column

- Change type of *rating_numerator* as decimal (som rating re like 13.5)
- Extract *rating_numerator* from tweet text

5) Clean *doggo, floofer, pupper, puppo* columns

- Extract dog stage from tweet text for *doggo, floofer, pupper, puppo* columns
- Convert all dog stages name to lowercase
- Combine *doggo, floofer, pupper, puppo* in one dog_stage column
- Drop column *text* as not needed any more

df_predictions_copy table

1) Columns to delete: *jpg_url,img_num*
2) Rename columns p1,p2,p3, p1_conf,p2_conf,p3_conf, p1_dog,p2_dog,p3_dog as not obvious column names
3) Convert all columns with dog breeds to lowercase
4) Choose only one most probable prediction with dog flag = True, drop other predictions columns
5) Drop all records where dog breed is not predicted (all dog flags = False)

1.0.2 Tidiness:

1) Combine cleaned *doggo, floofer, pupper, puppo* columns in one dog_stage column and drop columns *doggo, floofer, pupper, puppo* after that
2) Merge all three datasets into one: df_all_tweets Some tweets were deleted from Twitter after tweets were downloaded into archive file and some archive records were cleaned as invalid. Predictions dataset has only records for which dog breed was identified. Only records presented in all three tables will go into the merged table

Now I have only one dataset (df_all_tweets) where each record describes one tweet and has an image prediction for it.