wrangle_report

October 12, 2018

1 "We Rate Dogs" Data Wranglin Report

In this report, I show the flow of data wranging in detail.

1.1 Gather

I gathered these 3 datas needed and loaded in as a dataframe.

- twitter-archive-enhanced.csv
- image_predictions.tsv
- Additional archieve data (fav coount and retweet cdount) from Twitter API This data was saved in the local directory as "tweet_json.txt" and loaded into notebook as "df_api")

2 Assess

After assessing datasets visually and programatically, I found these problems which should be modified.

2.0.1 Quality

df_archive table

- "timestamp" is a string not a datetime.
- Rows of tweets which are later than 08/01 2017 should be removed.
- Tweets which are original ratings should be extracted.
- "name" column are unreliable, thus it should be delete.
- 'rating_numerator' and 'rating_denominator' columns are not necessarily correctly extracted.
- 'rating' column which represents ('rating_numerator / rating_denominator) should be created.
- Stage columns ('doggo', 'floofer','pupper', 'puppo') are not necessarily correctly extracted.

df_image table

- Some pictures are predicted not as dogs. If a picture is not predicted as dog till the 3rd prediction, delete that row.

2.0.2 Tidiness

- "df_api" and "df_image" should be merged to "df_archive"
- Dog stages in the archive data should be in 1 column.
- We need only the most primary confident prediction of dog types from pictures, so make the column "predicted dog type" in place of p1~p3 predictions.

2.1 Cleaning

I cleaned the data in this order.

- make copy of 3 datasets which would be modified
- "archieved_clean" table

    1. extract original tweets (not reply or retweet)
    2. delete "name" column
    3. convert "timestamp" datatype into datetime
    4. remove tweets later than 08/01 2017
    5. re-extract'rating_numerator' and 'rating_denominator' columns
    6. create 'rating' column from them
    7. make "stage" column which has categories 'doggo', 'floofer','pupper', 'puppo, 'blep' (as "archieve_clean2")

- "image_clean" table

    1. choose the most confidential dog_type from prediction 1~3, and make new column "dog_type" instead of them.
    2. "df_api" and "df_image" should be merged to "df_archive"

- inner-marge 3 dataframes and name new one "df"

3 Store

- store "df" in csv-file "twitter_archieve_master.csv"