We Rate Dogs Project
Wrangling report

Mohamed Makki

January 10th, 2020

This report is about data wrangling for datasets gathered by Twitter for rating dogs. The wrangling process have three procedures: **gathering**, **assessing** and **cleaning**.

1. Gathering:

The gathered datasets are from three different resources and formats:

- The first dataset (twitter-archive-enhanced) is in csv format, already manually downloaded; we have to read it in pandas to launch it for wrangling process (named as twitter_arch_df).
- The second dataset (image-predictions) is in tsv format, mainly collected by Udacity, this one are programmatically downloaded from Udacity servers (named as image_predict_df).
- The third dataset (tweet_json) is a json file in text format; this was downloaded using python's Tweepy library. (named as tweets_df).

These DataFrames are first visually inspected to get primary knowledge about data structure and main components of them, then will start manipulating data programmatically (by coding) to get further perception and grasp about the data.

2. Assessing:

In this stage of data wrangling, we estimate and define which part of data need our consideration for cleaning later on.

Here will do further sophisticated inspections and mining in our data, all assessing operations to find out these parts of data that may be needed to be cleaned and tidied up. Accordingly, we have the following issues.

Quality Issues

- In the first dataset (twitter_arch_df) some columns contains only 78 rows.
- Some other columns contains 181 rows.
- In the second dataset (image_predict_df), some dogs breed names should be capitalized.
- In denominator column some values != 10.
- In numerator column, some values are very high.
- In timestamp part of it should be removed . . meaningless information.
- The timestamp column should be in numerical format, not string format or object.
- Convert some columns from numeric to string.
- Changing column name when necessary.

Tidiness Issues:

- In the second dataset (image_predict_df) , there are four columns allocated for dogs breed names.

- In the above dataset, a new column have to be created in replacement of the dropped columns above.
- In the same above dataset, there are nine columns for dog prediction and degree of confidence (p1, p2, p3, p1_dog, p2_dog, p3_dog, p1_conf, p2_conf, p3_conf . . . should be reduced (into only two columns).
- In the same dataset, rating numerator & rating dominator need some adjustments.
- Instead of the above, a new column should be created; for dog_rating.

3. Cleaning:

First, we are going to create a copy of each dataset before start our cleaning process. This is necessary for maintaining our original datasets from loss during an irrecoverable faulty cleaning operation or irreversible non-thoughtful step. Afterwards the following cleaning steps have been made:

Quality Issues

- Delete two columns (in_reply_to_status_id) and (in_reply_to_user_id) in (copy_twitter_arch_df).
- Delete three columns (retweeted_status_id), (retweeted_status_user_id), and (retweeted_status_timestamp) in (copy_twitter_arch_df) which contains 181 rows.
- Capitalize dogs breed names in (copy_image_predict_df).
- Correct some values in rating_denominator columns as the values should all be constant = 10.
- Correct some values in rating_numerator which lie above the normal scale of dog rating as per rating terms.
- Delete part of timestamp value/string (+0000) in timestamp column with meaningless indication.
- convert 'tweet_id' in all datasets from numeric to string, note that the tweet_id in the third dataframe named 'id'
- Change column name from 'id' to 'tweet_id' and other column name from 'favorite_count' to 'likes' in the (copy_tweets_df).

Tidiness Issues

- Create new column "dog_stage" instead of four columns "doggo", "floofer", "pupper", "puppo" in (copy_twitter_arch_df) dataset; then drop the four columns.
- In the (copy_image_predict_df) we have 9 columns for prediction and confidence, we may create only two new columns one for dog prediction (prediction) and the other for confidence of this prediction (confidence).
- Drop the (rating_denominator) column (column with one value = 10).
- Rename the (rating_numerator) column to (rating).

After each cleaning operation, a code will applied then a test will be done to assure validity and exactitude of our coding.

Finally, after cleaning process will merge the cleaned dataset in one dataset (twitter_archive_master) in csv format.

The resulting dataset will used for analyzation and visualization