wrangle_report

January 19, 2019

In this project I had to gather, assess, clean and analyze a dataset of tweets about dogs, breeds, their names, scores and number of retweets and favorites. This dataset consisted of three data sets:

1. tweeter archive, provided by Udacity. I had to read that csv file by `pd.read_csv()`
2. tweeter dog's ranking and image prediction dataset. The url with the file was provided by Udacity, I had to gather it by using Python `request` library
3. To query Twitter Api using `tweepy` library to collect extended archive with data about number of retweets and favorites. This one I querried using list of `tweet_id` extracted from the first archive and the code, provided by Udacity, saved to `tweet_json.txt` and then read line by line into dataframe.

While doing assessing visually and programmatically I found:
Quality issues:

- all dataframes have different number of entries:

- twitter archive had 2356 rows

- rating and image prediction file had 2075 rows

- extended twitter archive had 2340 rows

Twitter archive:

- if the twitter archive many columns had missing data as: in_reply_to_status_id, in_reply_to_user_id, retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp, expanded_urls
- name looked like it has all values but visual check showed that there are many None (745 rows) or articles without names as an
- rating_denominator was not always 10, in fact it's not 10 in 23 cases
- rating_numerator was incorrectly picking up values had to be programatically extracted from `text` column
- the score had to be recalculated
- doggo, floofer, pupper,puppo columns shown 2356 values but many of them are None, and visual assessment shown we didn't have breed for most of entries (1975).
- 181 retweeted posts

Image prediction dataset:

- 559 cases when with probability of p1_conf less than 0.5 the picture was the picture of a dog
- in 543 cases in p1_dog there is totally different animal with hight p1_conf probability
- in 101 cases when p1_dog was False, but p2_dog and p3_dog were True, it was not a dog, so p1_dog False was correct
- in 324 cases there were False in all p1_dog, p2_dog, p3_dog and it was correct assessment of a photo

Extended Twitter archive(API)

- comparing to 2356 tweet_id in df_archive file we could get data only for 2340 ids. 16 ids failed.

- not all `tweet_id` matched through all datasets, in the final clean dataset I only merged `tweet_ids` which were present in all three dataframes by `inner` join.

Tydiness:

- doggo, floofer, pupper,puppo columns had to be converted into one column with dog's state instead of four. It also would make able to see missing states
- `tweet_id/ id` column was present in each set, had to be eliminated by merging tables on `tweet_id`
- tweet_id in Extended Archive was labeled as id
- tweet_id in first two datasets was integer instead of string

Suggestion:

- though we have 9 columns with breed and if it's a dog prediction it's very inconvenient and unclear to read, need column which says right away if it's dog and what is the bread
- timestamp needed to be converted to date and time.

Based on this assessment I defined how to solve these issues and cleaned all datasets. Then I've merged all three of them by `inner` join on `tweet_id` into one dataframe, saved to csv file `twitter_archive_master.csv` and analyzed it.