wrangle_report

May 20, 2019

1 data wrangling report of WeRateDogs Twitter data

1.1 introduction

The dataset that is wrangled here is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog

1.2 gathering data

- `image-predictions.tsv` was downloaded programmatically
- `twitter-archive-enhanced.csv` was downloaded manually
- data was gathered from `tweet_json.txt`

1.3 assessig data

1.3.1 `twitter_archive`

quality issues

- alot of `NaN` in `retweeted_status_id`, `retweeted_status_user_id`, `retweeted_status_timestamp`, `in_reply_to_status_id` and `in_reply_to_user_id` but that ok since it could be the the original status not a retweet or a reply
- `expanded_urls` column have some `NaN`
- `timestamp` and `retweeted_status_timestamp` are object
- `tweet_id`, `in_reply_to_status_id`, `in_reply_to_user_id`, `retweeted_status_id` and `retweeted_status_user_id` are numeric
- outliers in `rating_numerator` and `rating_denominator`
- there is hidden NaN in `name` column
- `expanded_urls` have duplicated urls

Tidiness issues

-

1.4 doggo, floofer ,pupperand puppo are nontidy *messy* and there is hidden NaN in them

1.4.1 image_predictions

- tweet_id is int
- p1 have some predictions that are not dogs some of them are not recognized dogs and some are not dogs
- p2_conf and p3_conf are exponential which is not very clear to compare
- some column name may not be very clear

Tidiness issues

- p1, p1_conf, p1_dog, p2, p2_conf, p2_dog, p3, p3_confand p3_dog are not tidy
-

1.5 image_predictions could be merged with twitter_archive

1.5.1 tweet_likes_retw

- tweet_id is int

Tidiness issues

- this dataset is separated from twitter_archive which make them messy

1.6 cleaning data

- a copy was made for each data and all cleaning was done to the copies
- image_predictions columns were renamed appropriatly befor merging it to twitter_archive
- all the data were merged to twitter_archive making it more tidy
- doggo, floofer ,pupperand puppo were converted to one column called stage making them tidy
- data types were setted appropriately
- all replies and retweets were removed and columns associated with them
- some wrong values of rating_numerator were corrected
- rating_out_of_10 column was created by that formula rating_numerator / rating_denominator * 10

1.7 more to be done

- collecting data about the actual dog type to analyze models sensitivity and specificity
- completing stage data

1.8 conclusion

- this was challenging project specially trying to load the json in txt file and reading data from it