wrangle_report

May 31, 2018

1 COLLECT

First, we read all three datasets into dataframe. The datasets are: - twitter_archive_enhanced.csv - manual download - image-predictions.tsv - downloaded by using request() - tweet_json.txt - can't access twitter, so I just downloaded the txt file.

2 ASSESS

By using visual assessment & programming assessments like info(), isnull() etc., we've found tidiness issues & quality issues for each of the dataframes.

2.1 In twitter_archive

2.1.1 Tidiness Issues

- stage should be the variable and doggo,floofer,pupper,puppo are values.

2.1.2 Quality Issues

- tweet_id should be object instead of int.
- Some dogs'names are 'None'.
- Some names are incorrect.(the/a/an)
- Some rating_denominators are not 10.
- Some dogs don't have any stages.
- Some records are from retweeted posts.

2.2 In tweet_data

None.

2.3 In image_predictions

2.3.1 Quality Issues

- tweet_id should be object instead of int.
- Some records does not have any picture.
- For some pictures, none of the predicitons are dogs.

2.4 For all three tables

2.4.1 Tidiness Issue

- all three tables should be merged together.

3 CLEAN

After gathering all the issues, we define our cleanning processes accordingly.

3.1 In twitter_archive_enhanced

- Change tweet_id from int to object.
- Merge columns doggo,floofer,pupper,puppo into one column named 'stage'.
- Delete records for retweeted posts.
- Re-extract names from text using regular expression. (pattern: H/here we have, T/this is, M/meet, name is, named + Name)
- Change all rating_denominators to 10, and change the rating_numerator accordingly. Delete those records without ratings.
- Split multiple urls of expanded_urls into multiple columns.

3.2 In tweet_data

None.

3.3 In image_predictions

- Change tweet_id from int to object.
- Create a column named 'type' which contains the predicted type of the dog. we use p1 as the value; if p1 is not dog, we use p2; else we use p3. If none of the predictions are dogs, we delete those records.

3.4 For all three tables

Merge all three tables into one with selected columns

- df1:'tweet_id', 'text', 'rating_numerator', 'right_name', 'stage'
- df2:'retweet_count', 'favorite_count'
- df3: 'jpg_url', 'type'

join on: df1.tweet_id == df2.tweet_id == df3.tweet_id

We first copy the dataframes and then do the cleanning on those copied ones. We first deal with completeness issue, then tidiness issues, and finally quality issues.

After cleanning, we save the dataframe into a single csv file 'twitter_archive_master.csv'.