wrangle_report

August 20, 2019

0.0.1 Data Wrangling Project Report: WeRateDogs (by Ugur URESIN)

Project Introduction: The dataset which is going to be wrangled 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. These ratings almost always have a denominator of 10. The numerators, though? Almost always greater than 10. WeRateDogs has over 4 million followers and has received international media coverage.

REQUIRED PYTHON LIBRARIES

- numpy
- pandas
- requests
- os
- tweepy
- json
- scipy
- seaborn
- matplotlib.pyplot

DATA GATHERING PROCESS 1. twitter-archive-enhanced.csv: The WeRateDogs Twitter archive. It's downloaded via the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/59a4e958_twitter-archive-enhanced/twitter-archive-enhanced.csv **2. image-predictions.tsv**: i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv **3. The Twitter API & JSON**: Each tweet's retweet count and favorite ("like") count at minimum, and any additional data you find interesting. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then read this .txt file line by line into a pandas DataFrame with (at minimum) tweet ID, retweet count, and favorite count. **Note that:** My twitter developer account application was not approved without a reason or any additional request. Thus, I imported the following file manually: File: tweet_json.txt URL: https://s3.amazonaws.com/video.udacity-data.com/topher/2018/November/5be5fb7d_tweet-json/tweet-json.txt

DATA ASSESSMENT PROCESS Given data is examined **visually** in Excel first. Then, some numerical assessments were done using some python libraries such as numpy and pandas.

Especially following methods are used: * info() * head() * tail() * sample() * describe()

And there had been found following data issues:

ATTENTION for the REVIEWER: My previous submission is commented as follows: "The fact that the rating numerators are greater than the denominators does not need to be cleaned. This unique rating system is a big part of the popularity of WeRateDogs. You can see this on Project Motivation -> Key Points section in the classroom." However, the aim is NOT to remove the rows that numerator > denominator! The aim is to remove the numerator that >>10 because they can be considered as **outliers**! (I assume this understood wrong due to my expression.) Thus, I kept this issue as it is. Thanks!

Quality Issues

1. Twitter-archieve dataframe contains rows in which the denominator is not 10. These rows should be removed. Why the denominator is 10? Because "they're good dogs Brent" (https://knowyourmeme.com/memes/theyre-good-dogs-brent")
2. The tweet-json dataframe has duplicates!
3. Twitter-archieve-enhanced dataframe contains rows in which the rating numerator is much bigger than 10. These rows should be removed.
4. Number of rows are not matched. 'Twitter_archive_enhanced' contains 2356 rows. 'image_predictions' contains 2075 rows. 'tweet-json' contains 2354 rows.
5. The type of 'tweet_id' should be object however it's int64 in twitter-archieve-enhanced and image-prediction dataframes.
6. Following columns in the "twitter-archieve-enhanced" dataframe contain NAN values in most. Thus, these columns will not provide information to us. These columns should be removed. *in_reply_to_status_id 78 non-null float64* in_reply_to_user_id 78 non-null float64 *retweeted_status_id 181 non-null float64* retweeted_status_user_id 181 non-null float64 *retweeted_status_timestamp 181 non-null object
7. In twitter-archieve-enhanced dataframe, the timestamp column has redundant numbers '+0000'.
8. In twitter-archieve-enhanced dataframe, the type of timestamp column is not datetime!

Structural Issues (Tidiness Issues)

1. There are 3 dataframes. However, only 1 data frame is needed.
2. Twitter-archieve-enhanced dataframe contains 4 columns: doggo, floofer, pupper, puppo. These columns are categories. Thus, these columns can be united in same column.
3. Image-predictions dataframe contains 3 predictions. Instead of 3 different prediction outcome, only the best can be used.

DATA CLEANING PROCESS Both quality-issues and structural-issues (given above) were fixed by using python libraries.

Quality Issue 1 Twitter-archieve-enhanced dataframe contains rows in which the denominator is not 10. These rows were removed because they may led to inproper results.

Quality Issue 2 The duplicates were dropped from the tweet-json dataframe. (In the process, this issue is intentionally left to the end because previous cleaning operations might yield another duplicates or simply might drop all duplicates).

Quality Issue 3 Twitter-archieve-enhanced dataframe contains rows in which the rating is much bigger than 10. These rows should be removed (QI-3). Because these values can be considered as **outliers** and they may led to inproper results. Thus, these rows were removed.

Quality Issue 4 Number of rows are not matched in the dataframes (QI-4). Redundant observations were dropped. (See also Quality Issue-6)

Quality Issue 5 The type of 'tweet_id' should be object however it's int64 in twitter-archieve-enhanced and image-prediction dataframes (QI-5).

Quality Issue 6 Following columns in the "twitter-archieve-enhanced" dataframe contain NAN values in most. Thus, these columns will not provide information to us. These columns were removed. *in_reply_to_status_id 78 non-null float64* in_reply_to_user_id 78 non-null float64 *retweeted_status_id 181 non-null float64* retweeted_status_user_id 181 non-null float64 *retweeted_status_timestamp 181 non-null object

Quality Issue 7 The timestamp column has redundant numbers '+0000' (QI-7). These numbers were removed.

Quality Issue 8 The type of timestamp column is converted to 'datetime' (QI-8).

Structural Issue 1 There are 3 dataframes. However, only 1 data frame is needed. Thus, these dataframes were merged based on tweet_id values.

Structural Issue 2 Twitter-archieve-enhanced dataframe contains 4 columns: doggo, floofer, pupper, puppo. These columns are categories. Thus, these columns can be united in same column (SI-2). And also, it should be considered that there are some rows in which there are **multiple stages** such as doggo-floofer instead of one!

Structural Issue 3 Image-predictions dataframe contains 3 predictions. Instead of 3 different prediction outcome, only the best can be used (SI-3). Based on the number of true predictions, there is no significant difference between the predictors. The best predictor has a prediction accuracy around 66% which is weak. Thus, the true predictions from the 3 predictors can be merged.

LIMITATIONS Given datasets contain missing information in some observations (tweets). Some of them were dropped from the datasets, the rest were used in 'Analyze' phase.

0.0.2 Conclusion

A master dataframe is cretad and checked both visually and using Python methods mentioned above. In the ANALYZE phase no problem is occured. Desired outputs and conclusion could be drawn from the master dataframe.

Reported by Ugur URESIN, MSc. Thanks!