wrangle_report

August 24, 2018

0.0.1 Project Goal

This wrangle report describes the process to create trustworthy data for WeRateDogs project analysis. We will go through topics that are necessary for data wrangling:

1. Gathering data
2. Assessing data
3. Cleaning data

0.0.2 Gathering Data

There are three parts for gathering data step:

1. The WeRateDogs Twitter archive **twitter-archive-enhanced.csv** This archive contains basic tweet data with 2,356 records and could be downloaded from the link: twitter-archive-enhanced.csv

2. The Image Predictions File **image_predictions.tsv** This file is hosted on Udacity's server and contains the dog breed predictions for tweet images using the neural network. The file could be downloaded programmatically using the Requests library in Jupyter Notebook with the URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

3. Additional tweet data through **Twitter API** 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 tweet ID, retweet count, and favorite count.

0.0.3 Assessing Data

After gathering all the data, we will asses them visually and programmatically for quality and tidiness issues and list them down for cleaning step.

1. Quality: We will idenify content issues (completeness, validity, accuracy, consistency) on each dataset

   `Twitter archive data`

   - Retweet data shouldn't be included

- Incorrect `rating_numerator` and `rating_denominator`
- **Doggo, floofer, pupper, puppo** columns should have NaN, not None when it is null
- Inaccurate dog **name**: such, quite, an, a
- Dog **name** should have NaN rather than None when it is null
- Timestamp column should be datetime type
- Missing **expanded_urls** (2,297 instead of 2,356)

`image_predictions.tsv`

- Missing records (2,075 instead of 2,356)
- **p1, p2, p3** should have consistent capitalization

`tweet_json.txt`

- Missing records (2,343 instead of 2,356)

2. Tidiness: We will focus on structure issues for all the dataset

- `image_predictions.tsv` should be part of the `Twitter archive data`
- Favorite count and retweet count in `tweet_json.txt` should be part of the `Twitter archive data`
- **retweeted_status_id, retweeted_status_user_id, retweeted_status_timestamp** should be removed since they are for retweet info
- **Doggo, floofer, pupper, puppo** columns should combine into one column **dog_stage**
- **Timestamp** could have extra columns **year, month, weekofDay** to analyze post trends in different time period

0.0.4 Cleaning Data

After identifying the issues for the all datasets, we can start to fix them. During this step, we will create copies for each dataset and manipulate mainly on these new copies. This will help us to keep original datasets and easily compare with the results. Sometimes, we also create temporary columns like `new_rating` and `new_rating1` for manipulation purpose. These temprorary columns along with the duplicate columns like `tweet_id` we got while combining datasets, should be removed in order to present our final result. For each issue, we will use Define, Code, and Test step to document the changes we make. Finally, we will store the clean tidy dataset with 1,991 records as **twitter_archive_master.csv**.