Wrangle Report
WeRateDogs Twitter data

by Alena Sukretna

Introduction

Goal: wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations. The Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for analyses and visualizations.

Data: The dataset that will be wrangled for analyzing and visualizing 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. The archived data contains basic tweet data (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets as they stood on August 1, 2017.

Wrangling data

There were a few main steps made during this data wrangling process.

1. Gathering data

Gather each of the three pieces of data:

1) The WeRateDogs Twitter archive (twitter_archive_enhanced.csv)

Download given file twitter_archive_enhanced.csv and set it as a dataframe.

2) The tweet image predictions (image_predictions.tsv)

Downloaded programmatically file image_predictions.tsv hosted on Udacity's servers.

This file contains the tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network.

3) 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 (tweet_json.txt)

Using Tweepy to query Twitter's API for additional data beyond the data included in the WeRateDogs Twitter archive. This additional data will include retweet count and favorite count.

The Twitter API is one that requires users to be authorized to use it. This means that before run API querying code, needed to set up own Twitter application. And before that, sign up for a Twitter account.

Create data frame tweets_df from the file tweet_json.txt with 3 the most important columns: 'tweet_id', 'retweets', 'favorites'.

2. Assessing data

After gathering each of the above pieces of data, assess them visually and programmatically for quality and tidiness issues. Detect and document at least eight (8) quality issues and two (2) tidiness issues.

After gathering we have 3 dataframes:
* df - WeRateDogs Twitter archive
* images - tweet image predictions
* tweets_df - each tweet's retweet count and favorite('like') count

Main quality Issues (Issues with the data's content)
- unapproriate dog names
- wrong data types
- missed data
- Tweets that has been retweeted should be removed

Main tidyness Issues (Issues with the data's structure)
- All data should be in 1 table for convenience
- Drop columns that are not needed
- Combine each dog stage column into a single column named "stage"

3. Cleaning data

It is where I fix the quality and tidiness issues that were identified in the assess step.

1) Copy dataframes
2) For each dataframe and each issue:
* define issue
* write code
* test

4. Reassess and Iterate

On a different stages of project sometimes I need to come back to one of above steps and do some changes or add something. This is natural working process.

5. Storing data

Store the clean DataFrame(s) in a CSV file with the main one named twitter_archive_master.csv. If additional files exist because multiple tables are required for tidiness, name these files appropriately.

Conclusions

Wrangling this data set was very interesting and challenging task for me. Especially using Twitter API to gather JSON data. Different tasks in cleaning part helped me to expand my knowledge. Final table that was received by join of three data frames seems convenient for futher analysis.