Wrangle and Analyze Data DAND Project

Submitted by Doug Olsen

In this paper we will describe our wrangling effort made in the section of wrangling WeRateDogs Twitter feed project.

Data wrangling consists of three steps as follows:

- Gathering data
- Assessing data
- Cleaning data

Gathering

Gathering Data for this Project composed from three sources of data as described below:

1. Twitter Archive File
2. Image Prediction File
3. Twitter API File

Gathering: Summary

Gathering is the first step in the data wrangling process. We could finish the high-level gathering process:

Obtained data as follows

1. Reading from csv file using pandas (twitter-archive-enhanced.csv)
2. Downloading a file from the internet (image-predictions.tsv) Downloading file using requests
3. Querying Twitter API (tweet_json.txt) Get JSON object of all the tweet_ids using Tweepy

Imported data into programming environment =(Jupyter Notebook)

I found the Twitter API part very time consuming and frustrating, i downloaded twitter-archive-enhanced.csv into microsoft excel before uploading the file into Jupyter notebook, this caused a bug in my program. Took 2 days to fix.Excel wass the root cause , when i downloaded and uploded the .csv file directly into Jupyter Notebook bug fixed.

For me the gathering phase was the most time consuming and difficult part of the project

Assessing

Assessment is done both visually and programmatically. Identifying quality and tidiness issues is the goal of the assessment.

Tidiness Issues

- Dog "stage" variable in four columns: doggo, floofer, pupper, puppo
- Join tweet_info and 'image_predictions' to 'twitter_archive'

Quality

- **Twitter Archive Data set:**
  - remove re-tweeted data.
  - rating_denominator should equal 10, there were many strange values.
  - rating_numerator <=20, there were many strange values.
  - timestamp is string , it should be converted to datetime object
  - remove rows with empty values = expanded urls (no images)
  - remove columns 'in_reply_to_status_id', 'in_reply_to_user_id'
  - remove rows with empty values and columns= retweeted_status_id,retweeted_status_user_id, retweeted_status_timestamp.
- **Image Prediction Data set**
  - images in the prediction model are not all dogs, only want to see dog predictions. Remove all non-dog image predictions.
- **Tweet Info (API) Data set**

Note: There were many additional tidiness and quality issues with these data sets

Cleaning

Cleaning has 3 steps as follows:

- Define
- Code
- Test

I removed a significant number of tweet_id observations in my cleaning process the original achive data had 2356 rows of data and my clean data had 1459 rows of data.

I decided to eliminate all ratings that did not make sense:

rating_denominator != 10

rating_numerator > 20

I also removed all observations that the image prediction algorithm did not identify the image as a dog.