Wrangle Report

~ BY SHUBH

Introduction

The purpose of this project is to put out my practice what I learned in data wrangling section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled is the tweet archive of Twitter user WeRateDogs. This WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog. This report would briefly describe my wrangling efforts.

Project details

The tasks for this project or for Data Wrangling are as follows:

• Gathering data

• Assessing data

• Cleaning data

Gathering data

The data for this project consist on three different dataset that were obtained as following:

• **Twitter archive file:** the twitter_archive_enhanced.csv was already provided by Udacity and downloaded manually.

• The tweet **image predictions**, i.e., what breed of is present in each tweet according to a neural network. This file image_predictions.tsv is hosted by Udacity's servers and was downloaded programmatically using the Requests library and URL information.

• **Twitter API & JSON:** by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and stored each tweet's in entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

While gathering this much data, I encountered lots of new code such as using requests library, os library, and tweepy and the other web scrapping using html which leads me to shiver for some time because all were new to me, but later on I become comfortable with them.

Assessing data

Once the three tables were obtained I assessed the data as following:

• Visually, I used one tool, i.e., Jupyter Notebook for printing the three entire dataframes separately.

• Programmatically, by using different methods (e.g. info, value_counts, sample, duplicated, groupby, etc).

Then I separated the issues encountered in quality issues and tidiness issues. Key points to keep in mind for this process was that original ratings with images were wanted.

Cleaning data

This part of the data wrangling was divided in three parts: Define, code and test the code. These three steps were on each of the issues described in the assess section.

First and very helpful step was to create a copy of the three original dataframes. I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original, so that we can control it's version.

Whenever I made a mistake, I could create another copy of the dataframes and continue working on the cleaning part.

There were a couple of cleaning steps that were very challenging. One of them was in the image prediction table. I had to create a 'nested if' inside a function in order to capture the first true prediction of the type of dog. The original table had three predictions and confidence levels. I filtered this into one column for dog type and one column for confidence level.

Other was cleaning a code that was to melt the dog stages in one column instead of four columns as original presented in twitter archive.

One very challenging cleaning step was when I had to correct some numerators that were actual decimals, this was headache. This issue was brought to my attention after the first Udacity review. Using Excel and visual assessment was not sufficient to verify those decimals. Therefore, I had to run a code in order to check those actual tweets (decimals numerators).

Conclusion

Data wrangling is a core skill for Data Analyst. I have used Python programming language and some of its packages. There are several advantages of this tool that is used by many data scientists.

• For gathering data there are several packages that help scraping data off the web, that help using APIs to collect data (Tweepy for Twitter) or to communicate with SQL databases.

• It is strong in dealing with large data sets.

• It can deal with a large variety of data (unstructured data like JSON (Tweets) or also structured data from ERP/SQL databases.

• It is easy to document each single step and if needed re-run each single step. Thus, one can leave a perfect audit trail (perfect for the accountant).

• One can re-run analysis automatically every period. Thus, we could actually re-run the dog analysis every month with much less effort now because I have set it up once.

• Handling, assessing, cleaning and visualizing of data is possible programmatically using code.