Data Wrangling Project

Part of submission for Project 2 in Udacity Data Analysis Nanodegree through Egypt FWD initiative

Performed on WeRateDog account tweets before 2 Aug. 2017

By George Atallah

Steps

1. Project Requirement analysis
2. Gathering Data
3. Analyzing Data
4. Cleaning Data

1- Project requirement analysis

The first step done was to clearly read the project requirements.

This information is very important because if we do not know the full and correct requirements we can do great work but it can also be in the wrong direction.

For the project at hand the following requirements were found

a. Data must be gathered from 3 different sources in 3 different formats
b. Data validity guideline (only original tweets, must have image)
c. At least 8 data quality and 2 data tidiness issues must be assessed and cleaned
d. Rating with numerator greater than denominator are normal
e. Clean data must be stored, analyzed and visualized
f. 2 reports must be submitted (this being one of them)

2- Gathering Data

The following data have been gathered

A. Downaloaded the file 'twitter_archive_enhanced.csv' from the project resources. This is readily given by udacity.
   In the notebook I created a new dataframe called "ta_df" and read this csv file directly into it
B. Downloaded the file 'image_predictions.tsv' from the supplied URL.
   In the notebook I created a new dataframe called "ip_df" and read this csv file directly into it
C. Collected tweets data from twitter through the twitter API based on tweet id previously collected from image prediction file.

These data were stored in a text file that contained jso objects for each tweet
In the notebook I created a new dataframe called “td_df” and read this text file to it line by line so to be able to avoid and issues with nested attributes and also to be able to copy only the attributes needed
After reading all the file and collecting the attributes in a dictionary this dictionary was converted into a dataframe.

3. Analyzing Data

Now there are 3 data frames, one for each se of data.

First a sample of each data was printed in the notebook. This is very helpful to take an idea of the data format and layout in each df.

Then from these samples predictions as to where data quality issues might arise
For example when observing the df ta_df it was observed that not all dogs had dog stages associated with them

Another example is image prediction in ip_df, many of them were not even dogs

Also in ta_df some dog names were not correct, many had ‘None’ others had regular English words as the tweet didn’t contain a name but had the expression ‘This is’ so when names were gathered the first word after this expression was treated as a name.

The process continued with a lot of visual assessment in jupyter and in excel, also many tweets were viewed on the web to better understand the pattern of data quality issues.

4. Cleaning the data

This is a straight forward step with little issues.

I began by tackling all data quality issues be it by removing rows or columns depending on the impact.
Then I changed some dataframe columns data types to be consistent
Then I merged the 3 dataframes together in one large data frame based on tweet_id
Ater this I dropped all columns that were no longer needed (e.g. retweeted_status, reply_status, etc.)
The this data frame was saved in a new csv file to be accessed anytime later.