12/03/2021

Report On Data Wrangling Steps:
Gather, Assess, and Clean

By : Ahmed Mohamed Soliman

Introduction:

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. The dataset that is wrangled 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. These ratings almost always have a denominator of 10.

Project details:

The tasks of this project are as follows:

• Gathering data

• Assessing data

• Cleaning data

Gathering data:

This project gathered data from the following sources:

• The WeRateDogs Twitter archive. The 'twitter-archive-enhanced.csv' file was provided to Udacity Students. This archive contains basic tweet data (tweet ID, timestamp, text, etc.) for all 6000+ of their tweets as they stood on August 1, 2017.

• The Tweet image prediction, i.e., what breed dogs (or another object, animal, etc.) is present in each tweet according to a neural network. This file was provided to Udacity Students.

• tweet_json.txt to gather each tweet's retweet count and favourite ("like") count at minimum and any additional data I find interesting ,because i found alot of trouble talking to twitter support about having thier API so i handeld the project with the tweet_json file.

Assessing data:

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

• Visually, I used two tools. One was by printing the three entire dataframes separate in Jupyter Notebook and two by checking the csv files in Excel.

• 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.

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 interesting cleaning code 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 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 that whoever handles data should be familiar with.

I have used Python programming language and some of its packages. There are several advantages of this tool (as compared to e.g. Excel) 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 (Twitter didn't help me too much with acquiring there API so that was a bit Challanging).

• It is strong in dealing with big data much better than Excel.

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