Wrangle Report

Amrut Deshpande

February 17th, 2018

Wrangle WeRateDogs twitter's data

Introduction

Data Wrangling is a widely used technique in the field of Data science and it's a very crucial part before any analysis is been conducted on the data set. It consists of three phases namely, Gathering, Assessing and Cleaning.

The wrangling effort comes into action when dealing with huge chunks of dirty real-time data. With extensive use of python's powerful library tools we wrangle the data and generate a master clean dataset which can be used for inferential statistics and effective analysis.

Gathering

In this phase, I gathered data from various sources. The twitter handle of WeRateDogs gave Udacity the access to their twitter archive in form of a CSV file and this twitter archive contains basic information pertaining to its tweets. Each tweet image was run by a convolutional neural network to predict their breeds and this prediction information was made available to us through Udacity's url in a form of a tsv file. This file was programmatically downloaded using python's request library.

Finally, using tweet ID, I queried twitter's API for each tweet's JSON data to extract additional meaningful information like favorite count, retweet count for effective analysis. I found this to be challenging as I did not have prior experience in querying APIs and also about JSON texts. But stack overflow was of huge help and my mentor Khushboo too guided me during the course. Must say, Udacity has a good mentoring process.

Assessing

After gathering all the sufficient data, next step was to assess it to identify quality and tidiness issues. Some of the quality and tidiness issues were addressed visually, but most of them were identified programmatically. Majorly,

- Retweets data information not relevant for analysis
- Dog stages has multiple variables as columns
- Source column is not human readable
- p1, p2, p3 columns starting letter of each word is not capitalized
- Extract correct Numerator rating and Denominator rating columns
- No int data types for all the ids present

I split the data frame sections and focussed on identifying quality issues first relevant to each data frame and then tidiness issues. This helped me gauge most of the information I could possible obtain from each data frame.

Cleaning

After assessing the data, we come to the crucial and exciting part of cleaning the data. I made sure I copied all the data before I began cleaning it. I followed the systematic procedure Define, Clean and Test procedure for each of the issues in cleaning process. Initially, I began addressing the retweet information, by deleting retweets and any variables associated with retweet as they are just duplicates of original tweet and do not generate any useful information.

- I tackled the datatype of each variable and converted it to their appropriate types, like converting timestamp datatype from object to datetime which will ease the calculations for this variable.
- I addressed the tidiness issue in tweet_data dataframe by melting various dog stages columns into a single column as these represent a single variable.
- I encountered a major issue in the numerator and denominator values. The ratings for the same were not exactly extracted from the tweet text. This could affect the analysis due to wrong rating data.
- I coded a regex to extract ratings from the tweet text and ensured the ratings are correctly extracted.
- I removed all the int data types of ids.

Additionally, it was necessary to split the timestamp column into date and time columns respectively. Utilizing pandas str.replace and str.capitalize i addressed the quality issues in images data frame. Json data frame id column was renamed to tweet_id to match the data's of different data frames before merging them to create a master dataset.

Conclusion

I am thankful to udacity for providing me an opportunity to wrangle the dataset and extremely pleased with handling the wrangling part by overcoming all the challenges. I am now ready and confident to pursue any wrangling challenges in future.