Wrangle_report

February 16, 2018

1 Wrangle WeRateDogs twitter's data

1.1 By, Mithun Muralidhar

1.2 Feb 2, 2018

1.3 Introduction

Data Wrangling consist of three stages : 1) Gathering 2) Assessing 3) Cleaning

As a Data Scientist/ Analyst, this 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 to generate a master clean dataset that is ready for effective analysis.

1.4 Gathering

In Gathering phase, we gather data from various sources. WeRateDogs gave Udacity, access to their twitter archive in form of a CSV file. This twitter archive contains basic information like tweet id, timestamp, text etc.. Each tweet image was run by a convolutional neural network to predict their breeds. This prediciton information was available to us through Udacity's url in a form of a tsv file. This file was progrmatically downloaded using python's request library. 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.

Gathering phase was a challenging one. I should thank my Mentor Mr. Dominic for his continuous assistance and support, whenever I was stuck in extracting the data. Stackover flow references suggested by Udacity were of great help in understanding the syntax and data extraction quering API's.

1.5 Assessing

After gathering all the sufficient data, next step was to assess it to identify quality and tidiness issuess. Some of the quality and tidiness issues were addressed visually, but most of them were identified progrmatically. They include but not limited to 1)Retweets data information not relevant for analysis 2)Rating numerator and denominator colmuns has outliers and incorrect values extracted 3)Dog stages has multiple variables as columns 4)Source column is not human readable 5)p1,p2,p3 columns starting letter of each word is not capitalized

I split the dataframe sections and focussed on identifying quality issues first relevant to each dataframe and then tidiniess issuess. This helped me guage most of the information I could possible obtain from each dataframe.

1.6 Cleaning

Having assessed the data, now comes the crucial and exciting part of cleaning the data. 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.

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. This was a challenging part and I learnt a great deal from it.

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

I am extremely pleased with handling the wrangling part by overcoming all the challenges. I feel ready and confident to pursue any wrangling challenges in future.