Data wrangle report

By Mahmoud Alaa Mitwaly

December 2020

As an assignment for the Udacity Data Analyst Nanodegree; This is report illustrates the main steps involved in the data wrangle of Twitter account “WeRateDogs”.

Data Gathering :-

In this step, collecting data takes place. For this project, there were three main sources for data to deal with:

1. Twitter_archive_enhanced.csv file, this file was delivered and downloaded manually to our working dictionary and then imported into our working environment using panda function “pd.read_csv()”.
2. Image_prediction.tsv is the second file that has been hosted on a webpage and downloaded from its relevant URL using the Requests library get function and pd.read_csv() pandas function. This file encompassed image predictions for the dogs breeds obtained through a neural network on most of the tweets in the archive file.
3. The final dataset was gathered from Twitter REST API via the Tweepy library by querying the API to obtain extra information pertinent to the tweets ids in the firs file , e.g. retweets count and favorite count aspects.

Data Assessing :-

In this step, we investigate our imported dataset both visually and programmatically for quality and tidiness issues.

1. The visual assessment done on spreadsheet application like excel and then the programmatic assessment is conducted in Jupiter notebook.
2. Quality issues addressed first then tidiness issue after it and at the end of each issue the order number of cleaning process.

Data Cleaning :-

I used my knowledge of python and searching over the internet i.e. google, stackoverflow, stackabuse etc for references and possible guideance to resolve the above mentioned issues to the best of my knowledge. There was lot trial and error for difficult cases where regular expressions had to be used but at the same time some things for instance dropping the not so useful columns was pretty straight forward.

Overall, I learned a lot about how to use python effectively and efficiently to clean data and store it.

Finally, once the data was ready I analyzed it using visualizations as document in act_report.pdf.