Udacity Project - Wrangle of Data
Twitter Data Sets
By Andrey Gorelov
2019

Introduction

This is the fourth project for the Udacity program. I am asked to take a look at the Twitter social media, specifically at the WeRateDogs twitter account. It is a page that rates (only positively) pictures of dogs that get sent to them from other twitter users. The objective is to be able to collect data from the internet and to perform the following three steps on the data sets collected:

- Work with data
    - Gather the data
    - Access the information inside the data
    - Clean the data
- Store the data and do the following:
    - 3 insights on the information
    - 1 visualization on the insight to show some type of visual analysis.
- Write a report about Wrangle Act and Analysis of the Data.

Gathering

In this part, I need to obtain three different data sets in order to complete this project. This is how I have obtained them:

- First project 'twitter_archive_enhanced' was provided to be downloaded from Udacity website. As I was following the project objective and motivation, it would come up on one of the pages. The file contained information like (timestamp, source, text, rating, name, etc)
- The second file of 'image_prediction' that consists mainly of animal information I get from the url link. This code and url itself was also provided by the Udacity Course.
- Last but not least, Twitter API and JSON. SInce I did not manage to get my Twitter Developer Account, I have downloaded both files, also provided by Udacity course. That data sets focused on dogs favourite and retweeted tweets.

Accessing

In this part of the project, I check that all the data sets acquired can display information, but more importantly this is the part where I focus on pinpointing Quality and Tidiness issues of the data sets. To properly pinpoint the issues there is a criteria to follow.

Quality:

- Completeness
- Validity
- Accuracy
- Consistency

Tidiness Rules:

1) Each variable forms a column
2) Each observation forms a row
3) Each type of observation unit forms a table

In my project, I found the following issues :

Quality:

1) Remove retweets columns and rows from the data. They are essentially duplicates of tweets

2) Further remove any columns and rows that are related to retweets

3) Remove 66 duplicated 'jpg_url' in the 'image_predictions' file

4) Timestamp is classified as an 'object' instead of 'datetime'

5) In archive, the Source column needs work as it is hard to understand the data

6) Incorrect values in rating numerators

7) Image_predictions has a column p1 that has '_' inbetween word. Can be removed

8) In archive, some of the dog names are 'None' or written with mistake (a, the, 0, etc)

Tidiness:

9) The 'stage' of the dog (puppo, pupper, floofer, doggo) can be modified.

10) Merge the the data sets because they are all related to Tweets variable (would make sense)

11) Remove columns that do not match in amount and if the column is not 'useful'

<u>Cleaning</u>

This is the most important part of the project. Here, I have to clean any data that did not fall under Quality or Tidiness criteria. While cleaning, a clear definition of the issue must be **defined**, then a desired **code solution** to be run to fix the issue, and finally **testing** of the expected clean data should take place.