Project 2 : WRANGLE and ANALYZE Data

The dataset that you will be wrangling (and analyzing and visualizing) 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. The numerators, though? Almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent." WeRateDogs has over 4 million followers and has received international media coverage.

Project Steps Overview

- Step 1: Gathering data
  - Archive data
  - Images data
  - Tweeter api data
- Step 2: Assessing data
  - Archive data
  - Images data
  - Tweeter api data
- Step 3: Cleaning data
  - Tidy issue
  - Quality issue
- Step 4: Storing data

1. Gathering data-frame

There are three data set that have been used in this project

a. The “WeRateDogs” Twitter archive. The file has been downloaded manually and was provided by udacity

Link : data/twitter_archive_enhanced.csv

b. The Tweet Image Predictions. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and the following URL:
https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image - predictions/image-predictions.tsv

c. Using the tweet IDs in the WeRateDogs Twitter archive, query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file.

2 Assessing data-frame

Summary of the findings

1. Data quality issue

    a. For df_arch

        1. the columns doggo, floofer, pupper, and puppo datatype is string "None" instead of NaN
        2. the column name has multiple stops words
        3. The ratings are incorrect both the numerator and dinomentor
        4. the data set contains retweeted data
        5. there are some replies in the data

    b. for df_image

        1.There are 2356 tweets in the the archive dataframe and 2075 rows in the images dataframe. This could mean that there is missing data, or that not all 2356 of the tweets had pictures.

        2.tweet_id is an integer and should be an object

        3.The columns "p1", "p2" and "p3" possess inconsistent data names

2. Tideness issue

a. For df_arch

        1. the columns doggo floofer pupper and puppo all mean the same thing and should be in one column
        2. the source column contains some tags(href)that needs to be removed
        3. the data sets are in separate data frames
        4. the column timestamp datatype need to be converted into date and time for it be manipulated on
        5. To many columns, there is need to drop unnecessary columns
        6. there are double links in the url columns

b. for df_image

        1. should be combined with df_arch
        2. the len of each document should be same

c. for Df_gat_data

        1. The len of each document should be same
        2. should be combined with df_arc

3 Cleaning the issues

    Tideness

define :

* Issue: the data sets are in different dataframes that contain the same information
* Solution: combine the archive data and the images into on dataframe
* Joining the datasets together to address the untidy nature of them being separate. Once they are joined together then I can address the dirty data issues and the remaining tidiness issues.
* combine the 3 data frames on the index key which is 'tweet_id

Define

To many columns, there is need to surpress unnecessary columns

Define: issue 6

cleaning the double URLS in the expanded_url column

Define

Issue no 2

The source column contains some tags(href)that needs to be removed

Cleaning the HTML residues present in the Source

Define

issue no 4

The column timestamp datatype needs to be converted into date and time for it be manipulated on

Code

Conversion of the date data type

Define

issue no 1

the columns doggo floofer pupper and puppo all mean the same thing and should be in one column

Solution: is to change create column that we show the dog stage

Quality Issue

define

issue no 1 The columns doggo, floofer, pupper, and puppo datatype is string "None" instead of NaN

Define

issue no (b3)

- The columns "p1", "p2" and "p3" possess inconsistency , also converting the name column in to lower cases

- Solution is The lower() method returns a string where all characters are lower case. Symbols and Numbers are ignored.

Define

issue no a2

- Multiple stop words, removing stop word will reduce the data set

Define

issue no a2 The ratings are incorrect both the numerator

- solution manually changing the numerator ratings and denominator ratings

4. Storing Data and Reports

The cleansed data is stored in 'twitter_archive_master.csv'.