A Report on Data Wrangling Efforts

Submitted By:
Navpreet Singh

Summary:

This is a brief report describing the data wrangling steps used to clean data for analysis. This report is divided into 3 sections describing efforts applied to Gather, Assess and Clean the data to make it ready for analysis.

Gather:

The main dataset which contains all the ratings, tweet details and urls are provided as part of this project.

Another dataset which contains the information about the breed of the dog and the image url was exists at udacity servers. I used **requests** library to connect to the provided url and download the dataset.

Still some important information was missing like likes (favorite) counts and retweet counts. But as I already had the id for each tweet present in the archived dataset. I used a library called **tweepy** to connect to twitter and download the required information in the form of **json** objects using the tweet id. After downloading all the required data in one text file, I have read the json objects using json library and created a 3rd dataframe from it.

Assess:

As a part of assessing the acquired data, I needed to find the issues present in the data to clean afterwards. For finding more issues, I have done visual assessment and programmatic assessment.

- For visual assessment, I checked for structural issues like there were 4 columns for specifying the dog stage.
- For programmatic assessment, I used pandas functions to check inconsistencies in the datasets.

The issues are related to **Quality** and **Tidiness** of data. The issue I found are given below:

➢ **Quality Issues:**

- The @WeRateDogs Twitter archive contains some retweets also. We only need to consider original tweets for this project since it was mentioned in the project description.
- *tweet_id* column having integer datatype in all the dataframes. Conversion to string required. Since, We are not going to do any mathematical operations with it.
- *rating_denominator* value should be 10 since we are giving rating out of 10.
- For some tweets, The values of *rating_numerator* column are very high, possibly outliers.
- *name* column has None string and a, an, the as values.
- Extract dog stages from the tweet text (if present) for null values.
- *timestamp* column is given as string. Convert it to date.
- Since we are not using retweets, *in_reply_to_status_id*, *in_reply_to_user_id*, *retweeted_status_id*, *retweeted_status_user_id*, *retweeted_status_timestamp* columns are not required.
- *image_df* contains tweets that do not belong to a dog.
- *source* column in main data frame is of no use in the analysis as it only tell us about the source of tweet.

➢ **Tidiness Issues:**

- A dog can have one stage at a time, still we are having 4 columns to store one piece of information.
- We are having 3 predicted breeds of dogs in the image prediction file. But only required the one with higher probability, given that it is a breed of dog.
- All 3 dataframes contains same tweet_id column, which we can use to merge them and use as one dataframe for our analysis.

Cleaning:

Cleaning is the major and most complex part of the data wrangling process. The whole cleaning process was divided into 3 parts i.e. Define, Code and Test.

- **Define** is where I specified the issue that I was going to fix along with some info regarding how will i fix it. .

- **Code** section will contain the actual code used to fix the issue with data. I have used mostly pandas for this task.
- **Test** section, to test whether the code changes has fixed the issue or not.

Using all these different steps of data wrangling, I had cleaned the data for further analysis and visualizations.