Wrangle Report

By

Mohamed Sobhy

Givens:

The project was about WeRateDogs twitter account that was doing a humorous job of rating dog in a unique style that led to a massive popularity to that account.

Data has been provided but from different sources.

1- An Enhanced Twitter Archive for old tweets with some messy data.
2- Image Predictions File from a neural network project that can classify breeds of dogs.
3- Additional Data via the Twitter API

Target:

Go through data wrangling process:

1- Gathering
2- Assessing
3- Cleaning

The final output from that process is build a new clean data set and save it to a csv file named “twitter_archive_master.csv”

I should also provide some insights about the new data set and add some visualizations

Process:

- I obtained my twitter API tokens to download JSON data for each tweet and save it all in one file “tweet_json.txt”.
- I put every data set in a separate pandas data frame and name them as follows:
    - twt_arc_en_df
    - json_tweet_df
    - img_pred_df

- I started to assess my data frames both visually and programmatically and have come to a list of quality and tidiness issues:

Quality

`twt_arc_enh_df`

- Hyperlinks for sources in `source` column.
- Retweeted rows not needed.
- Rows with no image not needed.
- Data type of `timestampe` column.
- Data type of `tweet_id` should be string.
- Some names in `name` column have words like(a, an, the, his, this ...etc)
- Missing tweets will have no json data.

`img_pred_df`

- Duplicated urls in `jpg_url` column.
- Data type of `tweet_id` should be string.
- Dog predictions have underscore separation.
- Dog predictions have small letters.

Tidiness

`twt_arc_enh_df`

- last four columns representing dog stage.

`img_pred_df`

- Dog prediction should be decided in one column.
- Confidence levels should be in one column.

- In the process of cleaning, I followed the procedure of
  1- **Define**
  2- **Code**
  3- **Test**

- I managed to clean all the points above and finally I was able to merge the 3 cleaned data sets to build the file “twitter_archive_master.csv”.

Notes:

- Sometimes I used Microsoft Excel for visually assessing data and then try to reach same results just to have the sense of numbers, dimensions and catch ideas.

- I used tableau to explore the final data set for some insights and I wish I had time to do some visualizations using python, jupyter notebook and pandas.