Wrangle Report

Introduction

The purpose of this project is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations by applying the knowledge we've learned from Udacity Data Analysis Nanodegree program. 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 Twitter archive is great, but it only contains very basic tweet information. Additional gathering, then assessing and cleaning is required for "Wow!"-worthy analyses and visualizations. This report briefly describes my rangling efforts.

Project details

The tasks of this project are as follows:

• Gathering data

• Assessing data

• Cleaning data

• Storing, Analyzing, and Visualizing data

Gathering data

The data for this project consists on three different datasets that were obtained as following:

• **Twitter archive file:** The WeRateDogs Twitter archive contains basic tweet data for all 5000+ of their tweets. It was provided by Udacity and downloaded manually.

• **The tweet image predictions**, i.e., what breed of is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and was downloaded programmatically using the Requests library and URL information

• **Twitter API & JSON:** by using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's Tweepy library and

stored each tweet's entire set of JSON data in a file called tweet_json.txt file. I read this .txt file line by line into a pandas dataframe with tweet ID, favorite count, retweet count, followers count, friends count, source, retweeted status and url.

Assessing data

Each piece of gathered data is assessed visually and programmatically.

• Visually, the three entire dataframes were displayed separately in Jupyter Notebook.

• Programmatically, the dataframes were assessed by using different methods (e.g. info, value_counts, sample, duplicated, groupby, etc).

Then I categorized the issues into quality issues and tidiness issues.

Cleaning data

Each issues identified in assessing section were cleaned one by one in three steps: Define, Code and Test.

Copies of the three original dataframes were created before the cleaning starts.

Functions were created to fix some major issues.

For twitter_archive**,** function “replace_names” was created to replace a/an/the in ‘name’ column with real dog names extracted from ‘text’ if available, otherwise set to ‘none’.

Ratings have been recaptured to accommodate decimal numerators. Incorrect rating_numerator and rating_denominator values were modified manually and programmatically. Functions were used to identify and update missing data for dog stage columns “pupper”, “doggo”, “floofer” and “puppo”. Finally, different dog stage columns were melt to form a single column “dogs_stage” to consolidate all the four columns.

For image_prediction dataframe, duplicated “jpg_url” values were identified and removed. Different image prediction results were compared and the first true predictions were captured by using function “image”. New columns “dog_type” and “confidence” were created to show the prediction results and the confidence level.

To achieve the tidiness of data, three dataframes were combined into one using tweet_id as the key.

Storing, Analyzing, and Visualizing data

The final clean dataframe is stored and saved in a CSV file named ‘twitter_archive_master.csv’. After analyzing the final dataframe, three insights were discussed and visualized.

Conclusion

Data wrangling is a core skill that whoever handles data should be familiar with.

Python and its packages have significant advantages in dealing with data as compared to other tools such as Excel:

- For gathering data there are several packages that help scraping data off the web, that help using APIs to collect data (Tweepy for Twitter) or to communicate with SQL databases. It allows us to collect data from various sources.
- It is strong in dealing with big data (much better than Excel).
- It can deal with a large variety of data (unstructured data like JSON (Tweets) or also structured data from ERP/SQL databases.
- It is easy to document each single step and if needed re-run each single step.
- One can re-run analysis automatically every period once set up.
- Handling, assessing, cleaning and visualizing of data can be done programmatically.