Data Wrangling Report

By Siddharth Shankar

Introduction

The project is primarily focused on data wrangling taught in the Udacity DAND program along with analyzing and visualizing 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.

The Data

Enhanced Twitter Archive

The WeRateDogs Twitter archive (*twitter_archive_enhanced.csv*) as provided by Udacity and downloaded manually contains basic tweet data for all 5000+ of their tweets, but not everything. One column the archive does contain though: each tweet's text, which is used to extract rating, dog name, and dog "stage" (i.e. doggo, floofer, pupper, and puppo) to make this Twitter archive "enhanced." Of the 5000+ tweets, it has been filtered for tweets with ratings only (there are 2356).

Additional Data via the Twitter API

Using the tweet IDs within the WeRateDogs Twitter archive, additional data is gathered for all 5000+ tweets via the Twitter API and Python's Tweepy library.

Image Predictions File

The tweet image predictions, i.e., what breed of dog (or other object, animal, etc.) is present in each tweet according to a neural network. This file (image_predictions.tsv) is hosted on Udacity's servers and is downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv

Project Details

The project tasks are as follows:

- **Gathering Data**: 3 different types of data (comma-separated (CSV), json text, & tab-separated (TSV)) via different means was loaded, fetched and downloaded in the jupyter workspace. Twitter_archive_enhanced.csv containing 5000+ tweets of WeRateDogs was provided by Udacity which was manually put and loaded into the jupyter notebook. The tweet IDs of those tweets was loaded into a variable for fetching additional data of the tweets using the Twitter API. During the process of fetching the data via Twitter API, slow internet / connectivity issue was encountered. The code for fetching the data via Twitter

API was improvised to re-try the failed download attempt once. The fetched data was stored in a text file named '*tweet_json.txt*' which was later read and additional information such as '*timestamp*', '*media_url_https*', '*favorite_count*', '*retweet_count*', '*user_id*', '*screen_name*', '*followers_count*', '*friends_count*', '*statuses_count*', '*verified*' were extracted from the *tweet_json* file. Afterwards, the image prediction file as provided by Udacity was downloaded via python's request library.

- **Assessing Data**: The downloaded data was loaded in a dataframe and visual assessments and programmatic assessments were carried out on all the 3 data. Several quality issues along with untidiness were encountered and noted down in the workspace.

- **Cleaning Data**: The quality and untidiness issues were resolved using the Define, Code, & Test approach. As a part of the process, the very first step was to create a copy of all the dataframes so that in case of any problem, the original data can be looked upon. Some of the cleaning was done with a single line of code, some were done using the loop so that repetition don't happen. At last, all the 3 cleaned data was merged into one, duplicate records were removed, certain cleaning was done and finally it was saved as a csv file into the programming directory as '*df_merged.csv*'.

- **Analysis & Visualization**: The cleaned data file was read into the dataframe and certain insights and visualizations were made which can be read in the '*act_report.pdf*' file.