Data Wrangle Report

By: Nagashri Nagaraj

Date: November 9, 2018

Introduction

The purpose of this project is to put in practice the skills learned in Data Wrangling section of Udacity Data Analysis Nano Degree Program. The dataset wrangled is from the twitter handle @dog_rates, more popularly known as WeRateDogs. WeRateDogs rate pictures of dogs sent to them with humorous comments. Account has more than 3.75 million followers with most popular post was of a dog marching in the 2017 Women's March, which was retweeted more than 50,000 times and favorited 134,000 times.

Brief description of wrangling effort:

Data wrangling is conducted with these 3 main steps:

1. Gathering Data
2. Assessing Data
    a. Visual Assessment
    b. Programmatic Assessment
3. Cleaning Data
    a. Issue
    b. Define
    c. Code
    d. Test

Gathering Data:

This project requires gathering data from 3 different sources:

a. Download the given csv file manually: twitter_archive_enhanced.csv
b. The image file for the 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 downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv. The dataset was downloaded using Requests library and URL information.

C. Using the tweet IDs in the WeRateDogs Twitter archive, queried 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. I had to include wait_on_rate_time=True , wait_on_rate_time_notify=True into the API constructor within the code so the connection would not timeout from the server. Each tweet's JSON data was written to its own line. Then read this .txt file line by line into a pandas DataFrame.

Assessing Data:

a. Visual Assessment – was conducted by printing entire data set on jupyter notebook and also checking csv file in Excel.
b. Programmatic Assessment – was conducted using different Python methods like value_counts, sample, duplicated, groupby etc.,

Listed a summary of quality and tidiness issues.

Cleaning Dataset:

A copy of all 3 assessed data set were made to avoid any unintended changes to the dataset during cleaning process.

Started the cleaning process with addressing issues like missing data, mislabeled information which was predominantly found in WeRateDogs Twitter archive. Also addressed the issues in prediction columns in image prediction dataset. Using pandas library str.replace()function replaced underscore between words with space and str.title() to capitalize.

Final step of data cleaning process was to inner join all 3 datasets on tweet_id to form a master data set containing all information needed for insight and data visualization. Pandas library pd.merg() was used for this task.

Conclusion:

Data wrangling is the process of cleaning, structuring and enriching raw data into a desired format for better decision making in less time.

This project was very challenging for me till date. Most time consuming for me was calling API. I learned a great deal regarding tweepy and calling API by doing this project.

Even though majority of dogs are rated over 10, there are some instances where the ratings are less than 5. I would love to do some digging on these for further analysis.