Data Wrangling Report

Introduction

Data Wrangling is the process of transforming and mapping data from “raw” data form into another format with the intent of making it more appropriate and valuable for a variety of downstream purposes such as analytics. This project is primarily focused on wrangling data using Python. I gathered data from three different sources. The archive contains basic tweet data. Each tweet image was run through a neural network . The predictions were downloaded programmatically queried data using Twitter API.

Data Gathering and Collecting

The first dataset named ‘twitter-archive-enhanced.csv’, which I downloaded manually and was the most easy to import. This was done by first importing Python's Pandas packages and loading the data into the ‘twitter_archive’ data frame, using the pandas ‘read_csv()’ function.
The next package dataset was downloaded from the URL provided by Udacity using the request package, and save the response to a variable called ‘file_name’ using a for loop after opening the file, using a iter_content method having the chunk size of 1024 . For the final dataset I first loaded the image prediction files and then web scrapping off twitter id’s using the ‘Tweepy API’ found in the documentation. I then make a new dataframe with three columns : tweet_id, favorites and retweets.I stored each tweet's entire set of JSON data in a file called tweet-json.txt. The time period needed to acquire the tweets id’s lasted almost 14 minute.

Data Assessing and Cleaning

After I imported the data needed programmatically and assessed it I found few issues related to data quality and data tidiness. Several columns had empty values and some columns have invalid names, that needed to be changed to a None value. The timestamp column is an object so it has to be a datetime object. I had to change the datatype of the date column to ’datetime’ and then put them into two new columns. In doggo, floofer, pupper, and puppo columns, null values are represented as "NaN" values. I had to create a new column named ‘dog_breed’ and to compact all the for columns in one. Source is in HTML format with a and \a tags surrounding the text. I change the values in the percentage columns from two tables

from proportions to percentages.I had to converst the tweet_id from all three columns to an integer datatype. After I made copies of all three tables I merged them into one and then proceeded to drop the unnecessary columns, to rename and sort them and to make a new column named ‘dog_breed’ with information extracted from doggo,floofer,pupper and puppo columns. All the final cleaned datasets where stored to csv files.

Project Details

The tweet archive of Twitter user @dog_rates, also known as 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. For this project, we took original ratings from WeRateDogs that have images. Not all of the original tweets in the dataset are dog ratings and few are retweets.

Project Tasks:

Data wrangling is consisted of: Gathering and Collecting data, Assessing Data, Cleaning ,Storing, Analyzing and Visualizing Data. The report of my data analyses and visualizations is in act_report.pdf.

Conclusion

Working at this project gave me a good understanding of data wrangling, using API and JSON data.I also understood importance of data wrangling steps before we actually go for visualizations.

Author

Andrei Tibuliac