

Wrangle Report

The Data Wrangling project was the hardest and most knowledge rewarding of all projects. I believe I learned what data engineers do (In the cycle of data science project). Being able to collect data from APIs, clean them and fetching SQL commands to merge are a great asset to learn!

First, there were three datasets to wrangle with. The first dataset is the archive of WeRateDog which is exclusive for Udacity students. This archive contains basic tweet information (tweet ID, timestamp, text, etc.) for all 5000+ of their tweets. Each tweet image was infused to a convolutional neural network to analyze the images of dogs and correctly identify their breeds. The CNN predictions were programmatically downloaded using the Requests Python library as a tsv file. Furthermore, using the tweet IDs from the WeRateDogs archive I queried each tweet's JSON data. Using Twitter API and Python's Tweepy library, I kept each tweet's entire set of JSON data in dictionary, which I would later use to analyze the tweet's retweet and like counts.

The data gathering was challenging, from getting twitter's approval to querying the Twitter API. When I have configured and gathered all the data, I started the assessment stage. In assessment, I ran quality check and some tidiness issues. The regular cleaning process by addressing missing data and inconsistent information, which existed in the WeRateDogs Twitter account archive. Then converted each column to it's original data format, changing the timestamp data into datetime objects, tweet_id from a number into a string object, and the rating columns into float objects.

The image predictions data also had a cleaning to do. For example, removing the underscore between the words. The final stage in the data cleaning process had to be inner join of all three datasets into a master csv containing all relevant information. All thanks to Pandas Libraries for being a great help.

In nutshell, this project was challenging even though I have dealt with rest API before but, I did not think it would be this hard with twitter API and JSON data. Though it was hectic to meet the requirements, I would love to do it again.

Meaad AlRshoud

February 2019