Wrangle Report

By Vincent Man
July 28, 2018

Introduction

The purpose of this project is to wrangle the Twitter data from @dog_rates (aka WeRateDogs) in order to find valuable insights. WeRateDogs rates people's dogs with humorous comments and rate the dogs with a denominator of 10 and the numerator can be as high as 1000.
To complete this project, I am required to gather, assess and clean the Twitter data.

Gathering Data

The project data consists of three different datasets:

- A csv file (**twitter_archive_enhanced.csv**) provided by Udacity.
- A tsv file (**image_predictions.tsv**) which is hosted on Udacity's servers and must be downloaded programmatically by using the Requests library.
- Twitter API and JSON – I used Tweepy library in Python to query the Twitter API for each tweet from the twitter archive file in order to extract the favourite and retweet counts for each tweet and then save this data as a JSON file called tweet_json.txt file using UTF-8 encoding.

Assessing Data

After gathering the required data from above successfully, I assessed the 3 datasets visually and programmatically in order to evaluate, look for any quality and tidiness issues and then start fixing them.

Visually I first used Microsoft Excel to get a good grasp of the Twitter data and then I printed the 3 tables in Jupyter Notebook.

Programmatically I used various methods to analyse the data such as value_counts, info, sample etc.

Quality issues include:

- Incorrect data type for timestamp and tweet ids
- Retweets and replies that are not needed for analysis
- Dog names that do not seem to be real names
- Incorrect format for ratings (numerator and denominator)
- Sources that contain html tags which can be confusing to read
- Entries (variables 'p1, 'p2' and 'p3') that do not start with a capital letter and contain underscores
- Variable headers that can be difficult to understand straight away

- Duplications of Tweet ids

Tidiness issues include:

- Dog stages entries were created as 4 separate variables rather than having only one variable for the 4 stages values
- Consolidate the 3 datasets into one

Cleaning Data

As for the final step in data wrangling, this is where I clean the data and solve the quality and tidiness issues that were identified in the Assessing data stage.
The process was divided into 3 parts:

- Define – specify the problem
- Code – solve the problem
- Test – check to see if it solved the problem or not

It is always best to create copies for the datasets in case of major errors and you can simply make another copy from the original data.

- The data type for Timestamp and tweet ids have been converted to the correct type
- Any observations that contained retweets and replies have now been removed
- Dog names that do not seem to be real names have been removed
- Changed the format for the rating variables to floats
- Replaced the Source's html tags with spaces
- Capitalised the first letter of each entry and replaced the underscores with spaces
- Renamed the confusing variables with clear and concise names
- Removed duplicated entries of one particular tweet id
- Melted the 4 variables (dog stages) into one variable
- Consolidated the 3 tables by using pandas merge

Conclusion

This project is by far the most challenging especially during the gathering data stage when I had to use various packages to perform web scaping and used Twitter API to collect the JSON data which was a long process.
However, I managed to complete the project and I really did enjoy it as I was able to deal with various kinds of data and fixed issues that I have never seen before.