Wrangle Report

This document describes the wrangling efforts gathering, assessing, and cleaning the WeRateDogs Twitter archive data.

By Ahmed Glal

Gathering the Data:

The data for this project was in *three* different formats:

1.Twitter Archive File — WeRateDogs: WeRateDogs downloaded their Twitter archive and shared it exclusively for use in this project

This was extracted programmatically by Udacity and provided as a csv file to use.

```
pd.read_csv('twitter-archive-enhanced.csv')
```

2. Image Prediction File: The tweet image predictions, i.e., what breed of dog is present in each tweet according to a neural network is stored in this file. It was hosted on Udacity's servers in tsv format and had to be downloaded programmatically using the *Url*.

3. Twitter API — JSON File: By using the tweet IDs in the WeRateDogs Twitter archive, I queried the Twitter API for each tweet's JSON data using Python's tweepy library.

Tweepy is an open source Python package that gives you a very convenient way to access the Twitter API with Python. You can find more details on *setting up an app in twitter* and *accessing Twitter API using Python*.

Assessing the Data

After gathering the data, the three tables were saved and assessed Visually and Programmatically. With both the assessments I looked for Unclean data in all the three DataFrames, i.e. for Tidiness and quality issues.

Quality: *Low quality data is commonly referred to as dirty data. Dirty data has issues with its content*. The Data Quality Dimensions are Completeness, Validity, Accuracy and Consistency.

Tidiness: *Untidy data is commonly referred to as "messy" data. Messy data has issues with its structure*. Tidy data is where:
1. Each variable forms a column.
2. Each observation forms a row.
3. Each type of observational unit forms a table.

Cleaning the Data

Cleaning means acting on the assessments we made to **improve quality and tidiness**.

Improving quality doesn't mean changing the data to make it say something different — that's data fraud. Quality improvement means *Correcting when inaccurate, Removing when irrelevant and Replacing when missing*.

Storing the Data

After cleaning the data, I found out that there was no need for three data sets. All the data could be easily converted into a single file. So, I concatenated the three DataFrames on a common attribute twitter_id, to create the twitter_archive_master.csv.

Analyzing the Data

Using this freshly cleaned WeRateDogs Twitter data, interesting and trustworthy analyses and visualizations were created to communicate back the findings.

WeRateDogs is a Twitter account that rates people's dogs with humorous comments. These ratings almost always have a denominator of 10. The numerators though is almost always greater than 10. 11/10, 12/10, 13/10, etc. Why? Because "they're good dogs Brent". WeRateDogs believes almost all dogs deserve a 10 or sometimes more than that. Each tweet's retweet count and favorite (i.e. "like") count are another interesting parameters.

Python libraries such as matplotlib and seaborn have helped me with some cool visualizations for my analysis.

Most Popular Dog Breeds

Analyzing my data based on the first predictions gave me the *top 10 most popular dog breeds*.

Retweets, Favorites and Ratings

Amazingly many of the tweets have been retweeted. Some of the tweets have nearly *40000* retweets.

I tried to find out if there is any relationship between the Retweets, the ratings, and the favorites through a Correlation plot, one of the coolest plots in python. I used the seaborn library to generate a heatmap.

Conclusion

I am thankful to udacity for providing me an opportunity to wrangle the dataset and extremely pleased with handling the wrangling part by overcoming all the challenges. I am now ready and confident to pursue any wrangling challenges in future

their data intimately and is always looking for ways to enrich the data. I have done the same using amazing Python Libraries. This project was one of my most fun data experiences, so much that I decided to write my first blog on it. You can view the project, along with my reports on my Github.