Wrangle Report

May 13, 2018

1 Project Wrangle and Analyze Data- Wrangle Report

1.0.1 *Author: Anthony T. O'Brien M.D.*

12th May 2018

2 Introduction

This is a document to review the wrangling process/effort during the realization of my wrangle and analyze project of WeRateDogs twitter database.

3 Context

The context of this project is to wrangle WeRateDogs Twitter data to create interesting and trustworthy analyses and visualizations.

This dataset is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

4 1. Gathering data breakdown

The good The process to authenticate the twitter API was straight forward, and only took a few minutes. The emphasise on security to not disclose any of the keys provided is critical, and a very important consideration by Udacity. At this point in the DAND importing data from .csv files is second nature and the URL import was very simple.

The bad Running tweepy api is time consuming, and this is where I think most people will get stuck. However, it is a limitation of the API and not much can be done about it. On the plus side, it gives enough time to consider other projects or plan ahead. My recommendation is once you are able to get a file from it, save the file into a separate .csv so you never have to run the API again.

The ugly The documentation to load and write the twitter API is terrible, and I believe this is where I had the most problems. In fact I know this is where I had the most problems. I spent about 5 days going over this, consulting multiple sources, until serendipidously I solved it. For some it is easy, for me it was a nightmare. The rest of the project I completed in a day. This was the challenge, there needs to be better documentation on this in the class for students who are less familiar with these processes.

5 2. Assesing data

The good I am confident assesing data, so this part was not too difficult, and it was a refreshing change from the grueling problem I had with the writing/loading of the twitter API. The approach to summarize the assesment into quality and tidy categories is very useful and an enjoyable way to assess the data. I do not have any bad or ugly reviews for this section.

6 3. Cleaning data

The good By using breaking down everything as done in the assesing data section, and then by using the define-code-test cycle, it was very easy to perform this section. In general all the cleaning methods were covered in other lessons in the Data Analysis Nanodegree, so here again it was enjoyable reviewing these techniques.

The bad The challenge in this section was converting the prediction of dog breeds and the accuracy of the neural network model into individual columns, however it was solvable after about an hour. Unlike the API load/write section which took days.

The ugly I did not experience an overly difficult process worth documenting.

7 4. The data analysis and visualization section

This was a simple section relying on what we previoulsy learnt in the DAND. Nothing negative to mention, however I did learn to use datetime in matplotlib which is a new skill.

8 5. Insights from the data

What I learnt about the data is the following:

Relationship between favorites and retweets

- 1) There was a linear relationship between the number of favorited tweets vs the number of retweets, although this relationship becomes less obvious over 30,000 favorites.

Importance of dog score

- 2) The impact of score on the relationship between favorited and retweeted tweet was negligable.
- 3) As time progressed the score went under regression to the mean and became less relevant as compared to the beginning. That * is to say that the scores tended to 1

Method of tweeting

- 4) The iphone was the most preferred method for perfoming tweets, followed by Web-Client and lastly tweetdeck

Most posted dogs

- 5) The top three dog breeds tweeted in this data set were the Golden Retriever, the Labrador and the Pembroke

Note To learn more about the process please view the wrangle act file or the act report

9 5. Conclusion

The project was pretty simple, the only difficulty I encountered was the twitter API read/write section. Afterwards the rest was enjoyable.