wrangle_report

June 22, 2019

1 My data Wrangling efforts for the project: wrangle_act

2 Process- Gathering:

We had to retrieve data from three sources, two of which were provided by the Udacity and one from the Twitter API. The first piece of twitter data provided as a .csv file was quite easy to just read it in my notebook using the python standard functions. The second piece of data, the image predictions was to be accessed using the file handler in python. This was made easy after quite a bit of learning via the associated lessons by Udacity on this topic. But it wasn't that bad. The third piece of data I gathered is the twitter data from the Twitter API. This was the most difficult part of the gathering process. For this last piece, I had to open a twitter account as I did not have one and then get the access credentials to be used later for actually retrieving the data. Once this was done, the most time consuming and at times frustrating part came where I wanted to put all this data which was in json format into a text file. To make my life easier, I studied the helpful links provided in the project descriptions to undertsand how json works and jsons helpful functions like dump() and load() to a text file. I spent a lot of time making mistakes during this process. Initially, I was stuck at the point where I want to get a status object from the API to write to a text file using json functions as it was giving me lot of errors. I had to do some research on what the status object is and found some help on stack overflow that helped me understand stand how to convert it to a python object. Instead of working on whole of the API data(which would require 30 min to run as a whole), I just did a code testing for just one tweet to see how it looked like and then decided to incorporate into the whole code. But this took a lot of time until I finally figured how to connect all the dots. I saw that it was not bad at all and was very much satisfied it was resolved finally. The second most difficult part was reading this text file line by line into a python list. This part also took some trial and error. Here I got to learn about how python is powerful for this purpose; again knowing about json functions helped a lot.

3 Process- Assessing:

This part was a little easier once I got all three pieces of data in place. First I looked at each data piece individually for quality issues and the tidyness issues in each. To do this I visually inspected each dataset by looking at multiple samples of 50 at a time to find any inconsistenices in each variable, column names, duplicate rows,formats,missing values etc. In the process, I found that visually inspecting the data using another application like google spreadsheet is very important as well. Then also programatically inspected each piece of data to look for null values, duplicates, wrong data types etc. The most work was to be done in the initial dataset that was gathered for

only the useful columns individually assessing each them as the next step in the assessing process. Then writing down issues with each piece of data to be helpful as a guide for the cleaning process.

4 Process- Cleaning:

Once the issues were assessed and written down, it was easier to work on it individually to fix each. This part invloved changing datatypes for some variables, renaming columns to make all three dataframes having consistentency for them to be combined later, creating new columns based on the existing data , dropping some columns, writing some functions to extract data and replace the erroneous values with the correct ones; thereby I was learning even more each time.This part of cleaning took the most time since it required fixes after each revision by my reviewer/mentor. After fixing issues in each dataframe I combined into a master dataset which I could then use for some analyses.

```
In [ ]:
```