Wrangle report

Introduction

The purpose of this project is to put in practice what I learned in data wrangling data section from Udacity Data Analysis Nanodegree program. Real-world data rarely comes clean. In this project using Python and its libraries, I will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it. This is called data wrangling. I will document the wrangling efforts in a Jupyter Notebook, plus showcase them through analyses and visualizations using Python (and its libraries).

The dataset that you will be wrangled, analyzed and visualized 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. 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. WeRateDogs has over 4 million followers and has received international media coverage.

Project details

In this project the following specifications are to be met.

- Gathering the weRatetweetDogs tweets data.
- Assessing the weRatetweetDogs tweets data.
- Cleaning the weRatetweetDogs tweets data.
- Storing, analyzing and visualizing the weRatetweetDogs tweets data.

Gathering data

The data in this project contains there different datasets that were obtained as follows.

- **Twitter archive file**: twitter_archive_enhanced.csv was provided by Udacity's and downloaded manually.
- **The image prediction file:** what breed of dog in present in each tweet according to neural network. This file – the image_predctions.csv - was hosted in Udacity's servers and downloaded programmatically using the request and URL provide by the Udacity.
- **Twitter API and JSON:** Originally I was supposed to query the twitter API using the tweet IDs in the twitter archive file. However, after four days of applying the twitter developer account and the twitter team asking me questions after question, I realized that I probably made mistake somewhere along the process and they were not going allow me an account. And since I was some days behind the deadline I opted the second not so good option. I download the twitter-api.py file and twitter-json.txt both provided by Udacity. After that I read the txt file line by line in to list which than I converted in pandas data frame.

Assessing data

Once the three tables were obtained I assessed the data as following:

- Visually, by first printing the heads, tails or samples of the data frames in jupyter notebook and second by checking the csv files in excel.
- Programmatically, by using different methods e.g. value_counts, info and etc. and then separated the issues into quality issues and tidiness issues.

Clearing the data

This part of the data wrangling was divided in three parts: Define, code and test the code. These three steps were on each of the issues described in the assess section.

First and very helpful step was to create a copy of the three original dataframes. I wrote the codes to manipulate the copies. If there was an error, I could create a new copy from the original. Whenever I made a mistake, I could create another copy of the dataframes and continue working on the cleaning part.

There were a couple of cleaning steps that were very challenging. One of them was in the image prediction table. I had to create a 'nested if' inside a function in order to capture the first true prediction of the type of dog. The original table had three predictions and confidence levels. I filtered this into one column for dog type and one column for confidence level.

Other interesting cleaning code was to melt the dog stages in one column instead of four columns as original presented in twitter archive.

Conclusion

Data wrangling is a core skill that whoever handles data should be familiar with. I have used Python programming language and some of its packages. There are several advantages of this tool (as compared to e.g. Excel) that is used by many data scientists (including the guys at Facebook).

- For gathering data there are several packages that help scraping data off the web, that help using APIs to collect data (Tweepy for Twitter) or to communicate with SQL databases.
- It is strong in dealing with big data (much better than Excel).

- It can deal with a large variety of data (unstructured data like JSON (Tweets) or also structured data from ERP/SQL databases.
- It is easy to document each single step and if needed re-run each single step. Thus, one can leave a perfect audit trail (perfect for the accountant).
- One can re-run analysis automatically every period. Thus, we could actually re-run the dog analysis every month with much less effort now because I have set it up once.