Wrangle Report

By Abderrahman BENYAHYA

26 September 2019

The Project 5 of the Nanodegree 'Data Analyst' Program of Udacity consist in gathering the data of WeRateDogs (a Twitter account that rates people's dogs with a humorous comment about the dog) from three sources: a given tsv file which includes the tweets, the ratings and the name of the dog, a JSON file constructed from the Twitter API and the programmatic download of a database predicting the class of the dog in the photograph. The project is aimed at applying the learned lesson in gathering, assessing and cleaning data as wrangling is one of the tasks of the Data Analyst. I worked from the classroom using the Jupyter Notebook and coding in Python.

1. The gathering of the data

The data that was given was not a problem as it is easy to import directly the tsv with the panda library. Nor the download programmatically using the code in Jupyter that was explained in the lesson.

However, the most difficult part was the coding of the import of data from API. Indeed, the getting of tweeter developer account was not a problem. I did not see the hint in the lesson that was very useful for the solution. I tried to import without taking into account the fact that it should have been line by line so the first file was not readable as JSON and I got error when I tried to store it as txt. So I tried to read it from the Notebook and extract directly the data and transform it into a DataFrame but I lost hours. Finally, I went into the student Hub and saw other students asking the same question and I followed the advice given by the tutor. But it took hours to read line by line and to have the data loaded in the environment as there were restrictions. As I have to reload the notebook and re-execute the kernel, I decided to convert the cell into a Markdown to save time.

For me, this part was the most difficult one as it was very knowing and hopefully I learn a lot.

2. Assessing the data

This part was very pleasant as it was an exploratory of the data at the same time. I tried to be very meticulous and I think a lost lot's of time because when I finished a part, I tried another code and discover other quality or tidiness problem. It was difficult because I had to adapt and change as there are consequence from a data to another one and I had the full picture only when I finished the work.

I tried to be very methodic and follow the learned lesson by doing it visually and programmatically.

3. Cleaning of the data

This part was the best one for me because I had to use imagination and all the tools given in the lesson to achieve the goal that I assign to myself: have a clean data. I came out with an error in the code (I transformed it into Markdown) that gave me another insight to clean an element that I did not see before when I tried to set the twee_id as integrator in the prediction table. I enjoy it but I regret I did not use loop more.