wrangle_report

April 18, 2019

1 Wrangle Report

This project consist of a demonstration of what it takes to wrangle a dataset : 1. Gathering data 2. Assessing data 3. Cleaning data 4. Storing 5. Analyzing, and Visualizing

1.1 Gathering

I was given two dataset : - twitter-archive-enhanced.csv containing an archive computed by Udacity. - An URL to retreive image-predictions.tsv, containing the dog's bread of some tweet

Then I need to use the Twitter API to gather more information on the tweet in twitter-archive-enhanced.csv. I just keep retweet_count and favorite_count but I could have used more info. And I better have to do it because as we will see the informations in the archive has many problems.

So Gathering was easy to do.

1.2 Assessing

Assessing consist of looking around the datas to find if they are good enougth to be used. And that's mostly not the case.

We have many think to check, manualy and programmaticaly. But it is worth the effort because if you don't do it all the work you will do after will be useless: bad data give bad insigths !

Pandas was a great help to compute some easy statistics (min, max, distinct values...) and visualize the data.

I have to be careful of what the data mean, to not consider some data as issue as they are part of We Rate Dogs history. Like the notation outliers.

I also check the Twitter documentation to understand what was expected.

1.3 Cleaning

The longest part, at least for me, was the cleaning.

Many errors has been corrected with the same process : pass a regex through the data to extract information more accuratly.

I spend lot of time building these regex to get more information.

An important step was to check that I've done revelant cleaning and note distroying informations.

1.4 Storing

It's an important part because you have to store the result of cleaning to be used by other people.

Storing data is easy, thanks to Pandas and SQLAlchemy.

1.5 Analyzing, and Visualizing

That's the funniest part, because you could, at last, show some insights. Sometime you will find new issues to clean, or another dataset to gather.

I spend time building a nice visualization of the best rated dog with MatPlotLib.

It was a lost of time in a scientific point of view, but making prety visualization is very important when you have to share your work.

It gives you document a better look and keep the public attention.

I will talk more about my cleaning in "act_report".