wrangle_report

November 28, 2021

1 wrangle_report

I conducted this Data wrangling project in 3 big steps.

1.1 First Step(Gathering)

In the first step I gathered data from various recouses in different ways.To be precise, - Firstly I downloaded data manualy(twitter_archive_enhanced.csv),It was provided by Udacity - Then I downloaded data programmaticaly (image_predictions.tsv) with help of modul called `requests`,there was link to download it.This data was about breeds of dogs (predicted by neural network) - Finally I used Tweepy to use API of Twitter to get data.First I got list of tweet_ids from (twitter_archive_enhanced.csv),then I used these ids to get JSONs after storing all of them in text (each in one line),I read this `txt` file line by line and appended data to list in order to make dataframe from it

The reason why I also downloaded programmaticaly is that when someone else is going to see my analysis,he or she can jus t open my Jupyter notebook re-run cells and when I try to conduct this analysis again but with updated data, I can also re-run these cells

1.2 Second step(Assessing)

In the second step,I assessed data both programmaticaly and visually.For Visial assessment,I used exteranl application`Excel` as it is not comfortable to use pandas for this case.For programmatic assessment,I used Pandas itself,methods `.info`,`.describe`, and other features like filtering.After finding issuse,I separated them into to groups Quality issue and Tidiness issue

1.3 Third step(Cleaning)

First thing I did in this step was getting copies of three gathered dataframes so that if I do mistake,It will not affect to orginal data.when ever I make mistake,I could re-copy dataframes and continue cleaning I divided cleaning each issue into three parts,define,code ,test. There was several challenging steps.First was in prediction table.There were 3 predictions of neural network(some of them was not breed of dogs).I got only one prediction which is breed of dog and with high confidence level.Another one was dog's stages in four columns,I used melt function to solve this issue

1.4 Conclusion

Data wrangling is important skill which every data analayst should have,as it takes much time and without it we can not do any analysis.Even If we do,I wil lead to wrong results