wrangle_report

August 11, 2022

0.1 Wrangle & Analyse: Project Report

0.1.1 Gathering Data

In this section of the project I was gathering the data. First I loaded the given dataset in local space. Then I fetched the predicting data with images from the provided url address. Lastly I was accessing data directly from twitter to save them in json file and in a new dataset called twitter_tweets. This was the most demanding part of the project, I proceeded at first with the given steps in the instructions through API authentification. I kept getting unknown errors and couldn't figure out why for a long time. Untill a mentor told me it is because twitter changed the levels of developer access and I would need an enhanced level to use this code.

Then I tried to apply for the enhanced access, and even though it was granted for couple days, the code still did not work, not sure why.

I was quite frustrated at this point and needed additional mentor help to figure out another approach. There were some posts in the knoweldge center, I replicated the approach but it still did not work and I kept getting uknown errors, until the mentors helped me after two days effort and long communication thread, to figure out the issues there.

To be honest this was super frustrating part of the project for me, took a long time and a lot of anger. I think it would be nice from you to update the additional resources part in the instructions to inform the people that this approach that is listed there no longer works and help students figure out the other approach because to be honest just after absolving the course if was not easy to figure this out. Without the mentors help I would not be able to do this.

0.1.2 Assessing Data

In the assessing data part I scanned the data both visually and programmatically for any possible data quality and tidiness issues. I could spot many but listed only the ones I decided to work on in the section below.

0.1.3 Cleaning Data

In this section I cleaned all the issues discovered and listed in the file. From data types, to removing columns, duplicates, joining columns into one, correcting the name column or keeping only the best prediction value, etc.

Then I merged the files into one master dataset and saved it into the workspace as a csv.

0.1.4 Visualisig Data

In the data visualisation part I was answering the following questions: 1. What is the highest number of retweets and which tweet is it?

2. What is the most common dog class?

3. What is the most common breed?