Report

We have 3 dataset to study: image_prediction, tweet_json and tweet_arhive

Wrangling data:

To retrieve the data, I first created a new folder in which I store the dataset. The datasets were read using the pandas library. I took a quick look at the dataframes and got to know the data they contained.

Access data:

At this level I started by actually examining each dataset. I started by noting the problems that each dataset contains. This involves finding the size of each dataset, paying attention to missing values, describing the dataset, etc. When examining datasets, attention should also be paid to small details such as column format, column names, etc. During this step, we already get an idea of how to solve the problem of the respective dataset. We enumerate the datasets that have duplicate rows.

Cleaning data:

To clean the data, we take into account the problems identified during the previous step. Misnamed columns can be renamed. It would already be necessary to enumerate the useful columns. To do this, remove unnecessary columns. It is necessary to check the format of each column and check if the format corresponds to the data of the column. Missing values must also be taken into account. To process them it is necessary to check if the missing values can be replaced or if they must be deleted. Both possible cases must be well thought out before being applied. It would be necessary to remove duplicate lines and be careful.

After these steps, you must start by analyzing the 3 dataset at the same time. We must see as far as possible if we can make a symbiosis of the three dataset. You have to find the columns that they have in common. And check the compatibility of these columns. After merging the dataset, it is still necessary to do some data cleaning work.

After that we can move on to data visualization to answer the questions asked.