Twitter Data Wrangling Report

February 4, 2019

`Wrangling Report`

Data wrangling, which consists of:

1- Gathering data 2- Assessing data 3- Cleaning data

This is dataset challenging me a lot because it is require me to gather the data from multiple sources including Twitter API , i went through the main process of data wrangling which are :

Data Input :

I collected data from diferent sources for analysis purpose, Udacity provide us twitter_archive_enhanced dataset as CSV file . I download it manually and importing it easily , then move the data into the relevant directory and import the data by using Pandas library (Pandas.read_csv) , this is dataset contain basic tweet data(twett.ID , text , name of dogs.....etc) The next dataset was tweet image prediction . I download it programmatically using Request Package , I took URL and sav it in response variable.Each tweet image was run through convolutional network to analyze the image of dogs and identify their breeds. Final dataset was the difficult one, I have to use Twitter API , The connection to Twitter API was done by using Tweepy , I registered as a client with Twitter , I created a new application and after a few days I received consumer token and secret. After that I created OAauthenticaion handler instance, and test the authentication , it run successfuly , I read and write Json library , the issue i faced is that using Json was difficult , I have to add wait_on_rate_limit = True and wait_on_rate_limit_notify=True to make sure that connection wouldn't time out from the server . After importing all necessary data , I jump into next step which is assessing data.

Data Assessment :

There are two types of assessment visually and programatically , I evaulated the dataframe looking for qulaity and tidness issues by using both visual assessment and programmatically assessment , visual assessment was not very accurate and to save time and effort i prefre programmatically assessment. In fact, If i dip deep into te dat i will need a lot of time to clean the data properly , and as it is requireed from udacity to detect at least 8 quality issues and 2 tidness issues. For visual assessment ,I already open the CSV files to check the data , I will exclude the columns that are not necassary in twitter archieve such as : source ,source,axpanded URLs ,doggo ,floofer ,pupper, puppo and text. In twitter API data set change column name from ID to twitter_ID to match the other data set. In Twitter Archieve dataframe i notice that in name column , there are many wrong names such as : the , such , my , O , not ,just , by ...etc with NON value , because it is impossible name is these values.In Tweet image dataframe I will replace "_" with space. For programmatic assessment , I will use .Info() to display concise summary of dataframe , display non values .Also, .dtypes to return the datatype of each columns. I discovered that imestamp should be datetime datatype ,and tweet_id should be string not int.Also, i used df_Twitter_API.dtypes to check the datatype for all variables , I decided to change Id into twitter_ID to match other datset

and ofcourse change the type into String , at the same time i will exclude unwanted columns. I summarized the issues into steps and in next step which is data cleaning .

Data cleaning :

During the cleaning I solved the problem step by step , some of these issues was to change the data type , others to change the missng values to NAN while the challenging task is to reomve the tweet that already retweet and remove tweet with no image. However, having defined step by step helped me a lot to focus on each point and at the same time to exclude unnecessary columns , which mea this step save my time a lot , I might not having it if i did not have a plan from the beginning , I am really excited and happy. The final step in data ceaning is to join all three tables by tweet_id , I used pd.merge function to meger all three tables together with inner join.

After cleaning i saved the file in one master sheet , the i analyze the data by asking some inquires and i tried to figure out the solution and relation between them , finally i plot the relashion between retweet_count and favorite_count.