wrangle_report

January 5, 2021

0.1 Wrangle Report

0.2 Introduction

In this lesson, we have learned numerous techniques to gather, assess and clean data programmatically. These skills are exceptionally useful since data in real-world is rarely comes clean. This is the work that needs to be done in the background before nice data visualization and machine learning models can be effectively produced.

In this project, we gathered data from different sources and in many formats, assess the data's quality and tidiness and programmatically clean it using common python data science packages (numpy, pandas, matplotlib, and seaborn).

The goal of this project is to create interesting and trushworthy analyses and visualizations from this Twitter archive. This will be accomplished by gather more data such as dog's image classification using convolutional neural network to predict dog breed and also additional data with Twitter API.

0.3 Project Details

1) Wrangle the data including (gather, assess and clean)

2) Store, analyze and visualize the data

3) Data Wrangling Report and Data Analyses/Visualization Report

0.4 Gather

First dataset was given in csv file format which is needed to load into the workspace directly. The second file (prediction of dog breed file based on neural network was hosted on Udacity's servers and is downloaded programmatically through requests library with the given url. The last retweet data was gathered through querying Twitter API through tweepy library. The data is then read as JSON file and read as .txt file. It was then loaded as csv file to the workspace.

0.5 Datasets

0.5.1 Enhanced Twitter Archive

The dataset is given in hand in csv format. This dataset is the tweet archive of Twitter user @dog_Rates, also known for WeRateDogs, which is a twitter account that rates peoples' dogs with humerous comment about the dog. The rating are out of 10. However, the rating can go beyond that, which is the unique feature for this account. This dataset contains basic tweet data (tweet ID, timestamp, text, and etc.) for over 5000+ tweets.

0.5.2 Image Predictions File

The file contains tweet image predictions of the dog's breed or other object using convolutional neural network. The file(image_prediction.tsv) is hosted on Udacity's servers. It was downloadded programmatically using the Requests library through the given URL.

0.5.3 Additional Data via the Twitter API

Each tweet's retweet count, favorite count, retweet date, retweet's user information were extracted from the Twitter API using Tweepy library. The data is stored as JSON and as txt file. The file is read to pandas Dataframe and joined with other datasets.

0.6 Quality Issues Needed to be Addressed (Assess)

df_twitter_enhanced dataframe

timestamp column needed to be datetime object instead of string

in_reply_to_status_id, in_reply_user_id, retweeted_status_id, retweeted_status_user_id should be int instead of float

datetime and timestamp columns are repeated

tweet text needed to be normalized in order to extract information from

gender column needed to be engineered in order to perform further analysis

tweet_length needed to be engineered in order to perform further analysis

source column needed to be clean to remove url information

df_image_pred dataframe

p1, p2, p3 needed to be merged to one as well as p1_conf, p2_conf, p3_conf

p1_dog, p2_dog, p3_dog are not needed, we only need one most confident dog breed prediction instead of 3 values

img number are not needed since the model prediction result is in the file

df_api_data dataframe

datetime column is string which should be converted to datetime

in_reply_to_user_id contains only one unique value, which does not explain any variability of the data, needed to be dropped

0.7 Tidiness Issues

3 datasets needed to be merged

date_time data from the additional API data and timestamp from archive data are repeated

in image prediction data p1,p2,p3 and p1_conf, p2_conf, p3_conf needed to be merged

0.8 Clean

The data is cleaned programmatically using python data science standard libraries (numpy, and pandas). Other libraries are used also such as nltk to process text data and sklearn to perform modeling. Of course, other sources on internet relating to how to use those libraries were searched during the project such as Stack Overflow, Google, Udacity Courses/Lesson for additional guidance. This experience had given me a real hand-on experienec as a data analyst to able to go from extract, transform and model/visualize the data. The visualizations and models were documented in act_report.html.