wrangle_report

February 2, 2019

1 Wrangle_Report

1.0.1 Introduction

Data wrangling is the process of cleaning and unifying messy and complex data sets for easy access and analysis and to be familiar with. In this project I will gather data from a variety of sources and in a variety of formats, assess its quality and tidiness, then clean it.The dataset that will be wrangling,analyzing and visualizing is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs.

Gathering Data process for this Project consists of three pieces of data which are:

1- The WeRateDogs Twitter archive file. which was downloaded manually by clicking the given twitter_archive_enhanced.csv link. 2- The tweet image predictions, This file (image_predictions.tsv) hosted on Udacity's servers and should be downloaded programmatically using the Requests library and the following URL: https://d17h27t6h515a5.cloudfront.net/topher/2017/August/599fd2ad_image-predictions/image-predictions.tsv 3- Each tweet's retweet count and favorite ("like") count at minimum. Using the tweet IDs in the WeRateDogs Twitter archive, I will query the Twitter API for each tweet's JSON data using Python's Tweepy library and store each tweet's entire set of JSON data in a file called tweet_json.txt file. Each tweet's JSON data should be written to its own line. Then I will read this .txt file line by line into a pandas DataFrame with tweet ID, retweet count, and favorite count. #### Gather step summary: Gathering is the first step of data wrangling process. Obtaining data from different resources. 1- Getting data from an existing file (twitter-archive-enhanced.csv), and Reading from csv file using pandas. 2- Downloading a file from the internet (image-predictions.tsv), and Downloading file using requests. 3- uerying twitter API (tweet_json.txt) Get JSON object of all the tweet_ids using Tweepy Importing that data into programming environment (Jupyter Notebook).

1.1 Assess

After finishing the first step which is gathering data, assess data will be the next step to asses them visually and programmatically for quality and tidiness issues. I will detect and document quality and tidiness issues.

Assess Summary:

Quality Completeness, validity, accuracy, consistency (content issues). Archive Dataset 1- timestamp column should be datetime with day, month and year. 2- Columns that will not be used for analysis must be deleted. 3- The name of some column have invalid names like 'None', 'a', 'an' it must be more clear. Images Dataset 1- There are missing values from images dataset. 2- Columns that will not be used for analysis must be deleted. 3- There are 66 images jpg_url duplicated they must be dropped. 4- Some tweets are have 2 different tweet_id one refer to the other. json_tweeets dataset 1- There is a tweet_id that duplicated 8 times. Tidiness Untidy data structural issues 1- Some columns in images dataset are not needed sauch as tweet_id and jpg_url. 2- All tables should be part of one dataset. 3- May be it is a good idea to add gender column in archives dataset.

Clean Cleaning data is the third step of data wrangling steps. It is to fix quality and tidiness issues that were identified in the assess step.