Project 3: Data Wrangling with

WERATEDOGS

Table of Contents

1. Introduction

a. Goal of project

In this project, we will gather data from a variety of sources and file forma ts and then assess data visually and programmatically for quality and tidiness. After assessing the datas, we will make it clean, merge them into 1 file and store the clean_data into a csv file. Not only so, we will process on our wrangled data to produce the insights of data.

- ✓ Some questions the insights of these datasets:
  - ✓ Question 1: Which is the most favourite name? Change in Name over time, relationship between name with rating_numerator, favorite_count, reteet_count?
  - ✓ Question 2: Stage and its relationship with tweet number, favorite_count and retweet
  - ✓ Question 3: favourite breed? Its relationship with rating_numerator, favorite_count, retweet_count
  - ✓ Question 4: Relationship between Favorite_Count, Retweet_Count and rating_numerator
  - ✓ Question 5: Which time are tweets posted or retweeted, or marked favorite?

b. Datasets overview

We have 3 datasets twitter-archive-enhanced.csv, image_predictions.tsv and tweet_json.txt with 3 different file formats (csv, tsv, json.txt) to read.

Data source

- ✓ Dataset 1: twitter-archive-enhanced.csv
  - Source: Download directly from the link given on Udacity
  - Method of gathering: Manual download to get file twitter-archive-enhanced.csv
- ✓ Dataset 2: image_predictions.tsv
  - Source: From the link given by Udacity (actually we can download it directly)
  - Method of gathering: Programmatical download via Requests to get file image_predictions.tsv
- ✓ Dataset 3: tweet_json.txt
  - Source: Twitter API
  - Method of gathering: using Tweepy with twitter_api.py from Udacity

2. Data Wrangling

a. Data Gathering

All the data will be downloaded like mentioned above into 3 DataFrame.
Then we will have a first look about the data and do assess the datas in the next phase

b. Data Assessing

At each data have some problems, we will try to find out problem in order to solve this problem at the cleaning data phase. Followings are the problems, which are listed with its datasets in order to be easy to follow and to do the cleaning as well.

Data Assessing

- Dataset 1: twitter-archive-enhanced-2.csv
  - **Tidiness issues**
    - ✓ Tidiness issues 1: Column headers are values, not variable names such as `doggo`, `floofer`, `pupper` and `puppo` should be stored in 1 columnn.
  - **Quality issues**
    - Quality issue 1: Some dog names incorrect, i.e. 'such', or 'actually',
    - Quality issue 2: Some rows `text` show that there are more than 1 dog mentioned. Some of their names are missing (Quality issue 2).
    - Quality issue 3: One tweet has no rating associated with it (rating numerators in text contain decimals).
    - Quality issue 4: Missing/incorrect `rating_numerator`.
    - Quality issue 5: `rating_numerator`for 2 dogs needs to be considered.
    - Quality issue 6: `rating_denominator` must be equal 10. Another value is not correct.
    - Quality issue 7: `rating_numerator` too high.
    - Quality issue 8: Column `timestamp` is not object type. It needs to be changed into datetime type.
- Dataset 2: image-predictions-3.tsv
  - **Tidiness issues**
    - Tidiness issues 2: 9 of the columns predicting the dog type will be reduced to `breed_prediction`. Because we do not tend to analyze weather people post a real dog or not, so we will not take columns of `conf` into consideration.
  - **Quality issues**
    - Quality issue 9: value in 'breed_prediction' mixing with _, -, lower case, upper casedog type can be reduced to 1
- Dataset 3: tweet-json.txt
  - **Tidiness issues**
    - Tidiness issues 3: **Join** with Dataset 1 & 2 on `tweet_id`d. Storing Data

After cleaning data and checking it carefully again, we will store it in to file csv.

3. Exploratory Data Analysis

With the exploratory Data Analysis, we try to analyze and give some conclusions about the data and its insights. With matplotlib, we can visualize the data better, and make the analysis more easily to understand.

4. Conclusion:

The data gives us a very interesting view about the dogs and the way people using twitter. At this project, we have a chance to collect data from web. And doing some challenging data cleaning, then making some interesting analysis. Fortunately, some question we want to know, we can find it from the data after wrangling.