UDACITY

Project Report

By Shikhar Sharma

DATA WRANGLING

Libraries Imported - numpy, pandas, tweepy, datetime, json, re, os.

Tools used- Jupyter notebook, MS-Excel.

Steps taken for the data wrangling process are as follows:-

1- DATA GATHERING PROCESS.

STEP-1 I set up my twitter application by making a twitter developer account in order to generate the Consumer API keys, and the Access Token and Access Token Secret. I used ‘Tweepy’ (An easy-to-use Python library for accessing the Twitter API) to query twitter API for each tweet's JSON data and store each tweet's entire set of JSON data in a file called **tweet_json.txt** file. I wrote each tweet's JSON data to its own line. Then I read this .txt file line by line into a pandas DataFrame with ‘tweet_id’, ‘retweet count’, ‘favourite count’, ‘display text range’, ‘lang’ and ‘created_at’ variables.

STEP-2 Downloaded the **twitter_archive_enhanced.csv** manually from resource section. This file contained ratings, text, dog type, urls and many more variables.

STEP-3 Downloaded **image_predictions.tsv**, a tab separated value file (**tsv**), and accessed it using pandas read_csv with sep='\t'. This file contained a table full of image predictions (the top three only) alongside each tweet ID, image URL, and the image number that corresponded to the most confident prediction (numbered 1 to 4 since tweets can have up to four images).

2- DATA ACCESSING PROCESS.

In data accessing process I assess data for two issues:

1- **Tidiness Issue**- This issue was related to the proper arrangement of data in the table for easier analysis. The requirements were as follows:

- Every variable forms a column.
- Every observation forms a row.
- Every observational unit forms a table.

2- Quality issue- This issue was related to the content for example incorrect data types, missing data, invalid data, inaccurate data, inconsistent data. I used mainly four data quality dimensions to access this issue. These were as follows:

- **Completeness** : This dimension is related with missing data. issue.

- **Validity**: This dimension is related with invalid data. For Example- negative weights, invalid datatypes, etc..

- **Accuracy**: In this dimension, although data is valid but issue is with its accuracy. For example- A person with weight 80 Kg and height 56 cm. Although, height 56 cm is valid, it is inaccurate to associate this height with weight 80 kg.

- **Consistency**: Finally, this dimension is related with the consistency in variables in a column. For example a state having its full names as well as its abbreviations(California-CA) in the same column.

I assessed quality and tidiness issues using two processes:

1- **Visual Assessment**- Used external software like MS-Excel. Apart from this, I used pandas functions like head(), tail(), sample().

2- **Programmatic Assessment**- Using pandas functions like info(), describe(), isnull(), duplicated().

3- DATA CLEANING PROCESS.

First, I cleaned the data for tidiness issues to make my job easy while doing cleaning for quality issues.

- **TIDINESS ISSUE** - I basically found 5 tidiness issues in my data. These have been discussed below:

twitter_c dataframe

1- ‘Text' contain non descriptive URLs and also ratings at the end of each text. It should be split into text, rating and url. But since ratings are already present, I will remove ratings from ‘Text'.

2- Making one column for features 'doggo','floofer','pupper','puppo' because column headers are values, not variable names.

3- Splitting source column and selecting only text instead of url and tags like‘<a’, '/a'.

tweets_c dataframe

1- Split day, month, date and year from created_at feature and make separate columns for each.

Forming master dataframes by merging on common variable

1- Merge 'twitter_clean', 'predictions_c', 'tweets_c' on tweet_id.

- **QUALITY ISSUE -** I found 11 quality issues in all three datasets in total. These have been discussed below:

twitter_c dataframe

1- Dropping the features which contain missing data like ‘in_reply_to_status_id', ‘in_reply_to_user_id'.

2- 'Timestamp' is an object datatype instead of datetime datatype.

3- Removing the data which include retweets in the form of 'retweeted_status_id','retweeted_status_user_id','retweeted_status_timestamp'. And than dropping these features as they contain missing data as well.

4- ‘Name' column contain invalid data which should be removed. But removing this data will also remove some important features data hence it will be better to convert all invalid names to NULL.

5- Removing ‘retweeted_status_id', ‘retweeted_status_user_id', 'retweeted_status_timestamp' variables because they do not contain any data.

predictions_c dataframe

1- Lowercase and uppercase 'p_1','p_2', and 'p_3' variables. Convert all of them to uppercase.

2- Remove tweets which do not have ratings for dog. I will take only those tweets for which neural network has predicted at least one of the top three predictions as dog 'TRUE'. If none of the predictions for dog is 'TRUE' than I will discard those tweets from the dataset since their is very low chance that the dog is present in the image.

3- Their are inconsistencies in p1_conf, p2_conf and p3_conf values. Some values are upto 6 decimal place whereas some are greater than 6. Take values upto 3 decimal place.

4- Change the datatype of feature 'img_num' from float to int.

tweets_c dataframe

1- Data type of 'Display_text_range' is object which should be converted into int by removing brackets and 0 which will be the same for every range.

2- Converting datatype of ‘date' and ‘year’ from object to int in twitter_merge data frame.

END.