A PROJECT REPORT

ON

PREPARATION ET ANALYSE DES DONNEES

Submitted to

UDACITY

In Partial Fulfilment of the Requirement for the Award of

DATA ANALYST CERTIFICATE
ALX-UDACITY

BY

GoldCat

AUGUST-2022

Contents

1 Introduction

In this project called Data Preparation and Analysis, we had to collect data from various sources and various formats, evaluate their quality and order, and then clean them using Python and its libraries. The dataset we processed is the archive of tweets from the Twitter user dogrates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

2 Steps

Our work was performed in 5 steps:

1. Step 1: data gathering
2. Step 2: data evaluation
3. Step 3: data cleaning
4. Step 4: data storage
5. Step 5: Analyze and visualize the data

3 Data gathering

In this step, we gathered 3 datasets.We gathered the Twitter archives of WeRateDogs under the dataset twitterarchivedf using the readcsv method of pandas. After that, we collected the image predictions of the tweets in tsv format using the library request. Subsequently, using the tweet IDs in the WeRateDogs Twitter archive, we queried the Twitter API to get the JSON data for each tweet using Python's Tweepy library and store the JSON data set for each tweet in a file called tweetjson.txt. From this tweetjson.txt, we extracted the number of retweet and favorite for each tweet in the Twitter archives of WeRateDogs.

4 Data evaluation

After collecting the three data elements, we made the evaluations visually and programmatically for the quality and tidiness problems.

5 Data cleaning

We detected and documented at least eight (8) quality problems and four (4) tidiness problems.

5.1 Quality Problems

1. Presence of retweets in the archive data frame which is of no interest to our analysis.
2. The data type of in_reply_to_user_id, in_reply_to_status_id are in float
3. The data type of in_reply_to_user_id, in_reply_to_status_id are in float
4. timestamp data type should be modify to date time
5. Record with rating_denominator less or greater than 10. This constitute an incoherence with reference to the rater documentation
6. Record with rating_numerator less than 10 or too high since these are probably outliers
7. Presence of HTML tags in source column.
8. Some names starting with lower case instead to start with capital case

5.2 Tidiness Problems

1. Since we decided to drop retweets, we should remove retweeted_status_id, retweeted_status_user_id and retweeted_status_timestamp columns from twitter_archive_df
2. The columns floofer, pupper, doggo and puppo should be merged into twitter_archive_df as stage columns and should have as data type category since their value is just limited to 4 possibilities
3. Join the columns retweet_count and favorite_count from df_status to twitter_archive_df for a better data nalysis
4. p1_conf,p2_conf,p3_conf can be used to evaluate how weel the deep learning prediction algorithm performs, so, so these columns should be added to twitter_archive_df

For tha analysis and visualisation, confere act_report.pdf

6 Conclusion

We were able to complete all of these steps in accordance with what we learned in the classroom in terms of data quality and tidiness. The most difficult step was querying the tweeter API using the Tweepy library. This took about 30 minutes to complete. In the future, we would like to explore other data collection techniques, such as web scraping.