WRANGLE REPORT

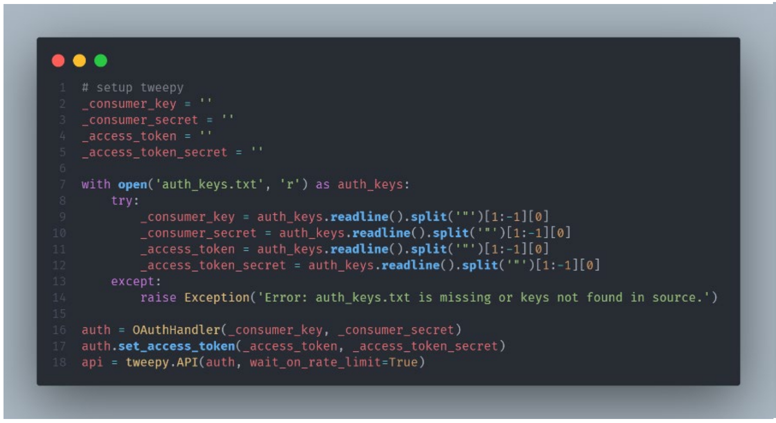
Bubuka Sharif

4th September, 2022

1. Gathering Data Phase

The datasets used were gathered from three sources. I manually downloaded the \`twitter_archive_enhanced.csv\` file, as it was openly made available by Udacity. I then programmatically downloaded the \`image_predictions.tsv\` file from the Udacity servers using the request library. The extra data about the tweets in the twitter_archive_enhanced.csv file was queried from Twitter through the Twitter API using a library called Tweepy. Each tweet's content was stored on a single line in a text file, and later read, to pick out the necessary fields for our project, and stored them in csv file.

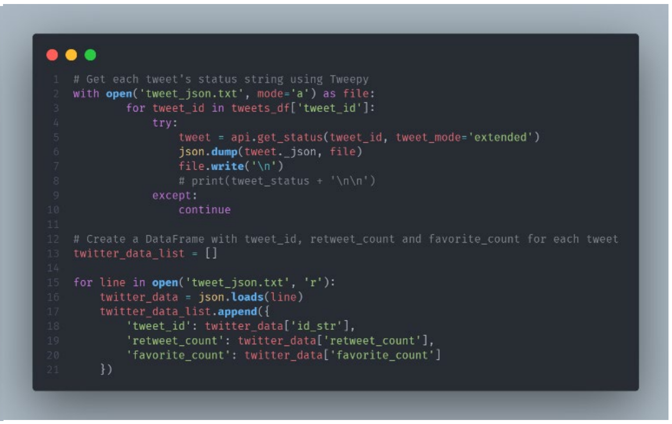
For clarity, here is a snapshot of the code snippet that I used to query the Twitter API using Tweepy in order to get the third dataset from Twitter.



```python
# setup tweepy
_consumer_key = ''
_consumer_secret = ''
_access_token = ''
_access_token_secret = ''

with open('auth_keys.txt', 'r') as auth_keys:
    try:
        _consumer_key = auth_keys.readline().split('"')[1:-1][0]
        _consumer_secret = auth_keys.readline().split('"')[1:-1][0]
        _access_token = auth_keys.readline().split('"')[1:-1][0]
        _access_token_secret = auth_keys.readline().split('"')[1:-1][0]
    except:
        raise Exception('Error: auth_keys.txt is missing or keys not found in source.')

auth = OAuthHandler(_consumer_key, _consumer_secret)
auth.set_access_token(_access_token, _access_token_secret)
api = tweepy.API(auth, wait_on_rate_limit=True)
```



```python
# Get each tweet's status string using Tweepy
with open('tweet_json.txt', mode='a') as file:
        for tweet_id in tweets_df['tweet_id']:
            try:
                tweet = api.get_status(tweet_id, tweet_mode='extended')
                json.dump(tweet._json, file)
                file.write('\n')
                # print(tweet_status + '\n\n')
            except:
                continue

# Create a DataFrame with tweet_id, retweet_count and favorite_count for each tweet
twitter_data_list = []

for line in open('tweet_json.txt', 'r'):
    twitter_data = json.loads(line)
    twitter_data_list.append({
        'tweet_id': twitter_data['id_str'],
        'retweet_count': twitter_data['retweet_count'],
        'favorite_count': twitter_data['favorite_count']
    })
```

2. Assessing Data Phase

2.1. Manual Assessement

I did some basic manual assessment of the datasets, not to really find out all quality and tidiness issues, but rather, to first get to know the layout of the tables and consistency in column values. I opened the csv files in Microsoft Excel to accomplish this.

2.2. Programmatic Assessement

Most of the assessment was done this way. Using various methods like .info(), .describe(), .sample(), .head() and .tail(), among others, I was able to assess various sections of the datasets for quality and tidiness issues. It is to note, that this was an iterative process with cleaning the data.

3. Cleaning Data Phase

3.1. Quality Issues

First, I looked out for data quality issues – that is issues to do with the content in the tables. Below is a brief outline of the issues I found, and my proposed solutions, respectively.

Issue	Source	Solution
Erroneous data types	All datasets	Change to suitable datatypes
Unclean (with HTML) values	Archive dataset	Remove HTML from value
Invalid dog names	Archive dataset	Replace invalid names
Inconsistent cases in names	Archive dataset	Capitalize all names

Wrong data	Archive dataset	Delete retweets

3.2. Tidiness Issues

After, I looked out for tidiness issues – that is issues to do with the structural layout of the data. Likewise, below is a brief outline of the issues I found, and my proposed solutions, respectively.

Issue	Source	Solution
Unmerged tables	All datasets	Merge tables into archive table
Many columns for dog stages	Archive dataset	Merge into one column
Two values in single column	Archive dataset	Break column into two

4. Storing Data Phase

With a single presumably clean master table, I was ready to start my analysis. Hence, I saved the table to twitter_archive_master.csv, as was requested in the project guidelines.