WRANGLE REPORT

Udacity Data Analysis Course

Remi Dongmo

Introduction

Real-world data rarely come clean. Using Python and its libraries, we will gather data from various sources and formats, assess its quality and tidiness, then clean it.

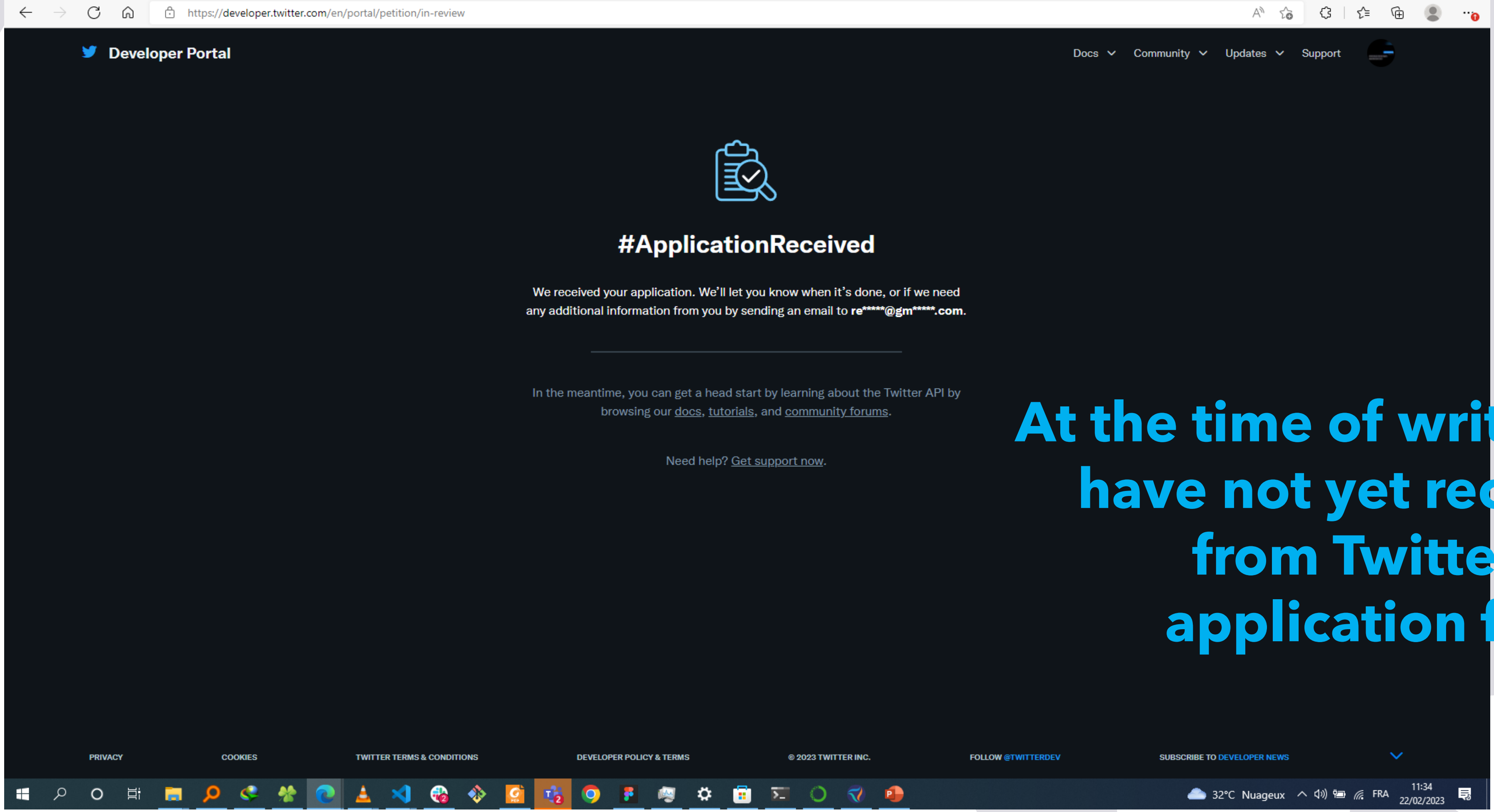
The dataset that we will be wrangling (and analyzing and visualizing) is the tweet archive of Twitter user @dog_rates, also known as WeRateDogs. WeRateDogs is a Twitter account that rates people's dogs with a humorous comment about the dog.

1. Data Gathering

Here we gathered data from three different sources:

- An archive of dog ratings in a csv file named: twitter_archive_enhanced.csv based on the WeRateDogs Twitter account activity
- A dataset named image-predictions.tsv containing predictions of rated dogs' breeds that were to be downloaded online using the python requests library
- A json file named tweet_json.txt planned to be obtained from fetching the actual tweets data on the twitter account page using the tweepy library. Unfortunately, since the Twitter team haven't yet even responded to my Developer account application (Sent more than 4 days ago) I had to manually download the json text

1. Data Gathering



#ApplicationReceived

We received your application. We'll let you know when it's done, or if we need any additional information from you by sending an email to **re*****@gm*****.com**.

In the meantime, you can get a head start by learning about the Twitter API by browsing our docs, tutorials, and community forums.

Need help? Get support now.

At the time of writing this report, I have not yet receive a feedback from Twitter concerning my application for a ddeveloper account

2. Data Assessment

During the assessment we could observe several issues within our datasets. We organized them into 2 categories:

A. **Tidiness issues**

1. The following columns in the archive dataset: doggo, floofer, pupper and puppo had the same nature but were separated
2. All the datasets were related, but separated as well
3. Some column names in the predictions dataset (*p1*, *p1_conf*etc..) were not very intuitive
4. On the predictions dataset only one prediction was necessary for analysis
5. Most columns of the twitter API dataset would not be necessary, we would only need 3 of them

2. Data Assessment

B. Quality Issues

1. The text column from archive dataset shows that decimal ratings where not well registered on the rating_numerator column
2. the rating_numerator column on the archive dataset has some values far superior to 10 (100+) from rating more than one dog on a same tweet
3. the rating_denominator column on the archive dataset has some 0 and less than 10 values
4. the timestamp and retweeted_status_timestamp columns on the archive dataset are of type String (object) instead of timestamp/date
5. Some fields on the archive dataset (name and also dog category fields) have invalid values like None, a, etc..
6. Based on in_reply_to_status_id there are 78 tweets that are replies and that we dont need

2. Data Assessment

7. Based on retweeted_status_id there are 181 tweets that are retweets and that we don't need
8. The source column is an HTML tag and not very suitable for analysis
9. The predictions dataset has fewer entries than the archive dataset, hence some images are missing
10. The Tweeter API dataset has 2 entries less than the archive dataset
11. The id column in the Twitter API dataset should be renamed as tweet_id to match the other datasets

3. Data Cleaning

During our cleaning process, we used a Define-Code-Test approach for every issue. At first, we made a copy of our assessed datasets, then we used programmatical methods to do the cleaning. Some important cleaning steps were:

- Fixing rating numerators and denominators by assigning correct values using RegEx and averaging all the ratings to make them over 10.
- Replacing all the « None » strings in the dataset by NaNs
- Removing tweets that were retweets or replies based on relevant variables
- Selecting the most accurate dog predictions while ensuring that the image predicted contained a dog

3. Data Cleaning

- Merging all dog stages columns into a single column
- Removing all invalid dog names obtained both through programmatical assessment and manual assessment on google sheets
- Removing HTML tags from the source column using Beautiful Soup
- Removing unnecessary columns from the Twitter API dataset.
- Finally merging all the datasets into a single dataset