Tripadvisor Reviews - Oberoi Delhi

Contents

This some project....!

Check out the content pages bundled with this sample book to see more.

Web Scraping Review Data From Trip Advisor For Oberoi Delhi

```
import requests
from bs4 import BeautifulSoup
import pandas as pd
from tqdm import tqdm
```

```
# Set up requests for trip advisor
headers = {
        'Access-Control-Allow-Origin': '*',
        'Access-Control-Allow-Methods': 'GET',
        'Access-Control-Allow-Headers': 'Content-Type',
        'accept': '*/*',
        'accept-encoding': 'gzip, deflate',
        'accept-language': 'en,mr;q=0.9',
        'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64)
AppleWebKit/537.36 (KHTML, like Gecko) Chrome/95.0.4638.69 Safari/537.36'}

url = "https://www.tripadvisor.ca/Hotel_Review-g304551-d304216-Reviews-
The_Oberoi_New_Delhi-New_Delhi_National_Capital_Territory_of_Delhi.html"
req = requests.get(url,headers=headers,timeout=5,verify=False)
print (req.status_code)
soup = BeautifulSoup(req.content, 'html.parser')
```

```
/Users/aalokatre/.pyenv/versions/3.9.2/lib/python3.9/site-
packages/urllib3/connectionpool.py:1045: InsecureRequestWarning: Unverified HTTPS
request is being made to host 'www.tripadvisor.ca'. Adding certificate
verification is strongly advised. See:
https://urllib3.readthedocs.io/en/1.26.x/advanced-usage.html#ssl-warnings
  warnings.warn(
```

```
200
```

```
# Loop through review pages and create a list of list for reviews
master_review_list = []
# for complete review set 0, 2900, 10
# for last page 2880, 2900, 10

for x in tqdm(range(0, 2900, 10)):
  url = "https://www.tripadvisor.ca/Hotel_Review-g304551-d304216-Reviews-or" +
str(x) + "-The_Oberoi_New_Delhi-
New_Delhi_National_Capital_Territory_of_Delhi.html#REVIEWS"
  # url = "https://www.tripadvisor.ca/Hotel_Review-g304551-d304216-Reviews-or2890-
The_Oberoi_New_Delhi-New_Delhi_National_Capital_Territory_of_Delhi.html#REVIEWS"
  req = requests.get(url,headers=headers,timeout=5,verify=True)
  soup = BeautifulSoup(req.content, 'html.parser')
  for y in soup.body.find_all(class_="YibKl"):
    review_content = []
    # Review content
    if y.find("q", {"class": "QewHA"}) :
      review_content.append(y.select_one('q[class*="QewHA"]').text)
    else :
      review_content.append(None)
    # Trip type i.e. solo, family, couple, business - we will work with this if we
have the time
    # if y.find("span", {"class": "TDKzw"}) :
    #   review_content.append(y.select_one('span[class*="TDKzw"]').text)
    # else :
    #   review_content.append(None)
    # Date of stay
    if y.find("span", {"class": "teHYY"}) :
      review_content.append(y.select_one('span[class*="teHYY"]').text)
    else :
      review_content.append(None)
    # Rating
    if y.find("div", {"class": "Hlmiy F1"}) :
      review_content.append(y.find("div", {"class": "Hlmiy F1"}).span['class'][1])
    else :
      review_content.append(None)
    # Owner's response
    if y.find("span", {"class": "MInAm"}) :
      review_content.append(y.select_one('span[class*="MInAm"]').text)
    else :
      review_content.append(None)
    master_review_list.append(review_content)
```

```
100%|██████████| 290/290 [08:39<00:00,  1.79s/it]
```

```
# Convert list to a data frame
reviews_df = pd.DataFrame(master_review_list, columns = ['Customer Review', 'Date
Of Stay', 'Customer Rating', 'Owner Responded'])
```

```
# Write scrapped data to csv file
reviews_df.to_csv('../data/oberoi_delhi_reviews.csv', index = False)
```

Data Pre Processing And Cleaning

Apart from customer reviews we have columns like ‘Date Of Stay’, ‘Customer Rating’, and ‘Owner Responded’. This data is scrapped from the tripadvisor website as is and need some pre procesing / cleaning to be in a usable format

```
import pandas as pd
import numpy as np
import re
from datetime import datetime
```

```
df = pd.read_csv('../data/oberoi_delhi_reviews.csv')
df.head()
```

	Customer Review	Date Of Stay	Customer Rating	Owner Responded
0	Excellent and highly recommended! We only sp...	Date of stay: November 2022	bubble_50	NaN
1	This was our favourite hotel whilst we visited...	Date of stay: August 2022	bubble_50	Dear Guest, Thank you for choosing to stay wi...
2	We stayed 2 nights at the New Delhi Oberoi. Wh...	Date of stay: November 2022	bubble_50	NaN
3	The Very BEST!! Quality of service and food is...	Date of stay: December 2022	bubble_50	Dear Guest, I am delighted you had a memorab...
4	The service , food, location, cleanliness is e...	Date of stay: December 2022	bubble_50	Dear Guest, I am delighted you had a memorab...

```
# Custom function to set trip type value
def set_customer_rating(customer_rating) :
  if (customer_rating is np.NAN) :
    return
  elif 'bubble_50' == customer_rating:
    return 'Excellent'
  elif 'bubble_40' == customer_rating:
    return 'Very Good'
  elif 'bubble_30' == customer_rating:
    return 'Average'
  elif 'bubble_20' == customer_rating:
    return 'Poor'
  elif 'bubble_10' == customer_rating:
    return 'Terrible'
```

```
df['Customer Rating'] = df['Customer Rating'].apply(set_customer_rating)
```

```
# Set owner's response from collected date (Yes or No)
def set_owners_response(owners_response) :
  if (owners_response is np.NaN) :
    return 'No'
  else :
    return 'Yes'
```

```
df['Owner Responded'] = df['Owner Responded'].apply(set_owners_response)
```

The extracted date is in the form of string. We will convert this string into date format, so that it becomes easier to perform EDA using this column

```
# Extract date from string
import datetime
def set_review_date(date_string) :
  if (date_string is np.NaN) :
    return
  else :
    extracted_date = date_string.partition(': ')[2]
    return datetime.datetime.strptime(extracted_date, '%B %Y').strftime('%m/%y')
```

```
df['Date Of Stay'] = df['Date Of Stay'].apply(set_review_date)
```

Write the cleaned data to a csv file. The output file will have reviews that are yet to be cleaned which will do further

```
df.to_csv('../data/cleaned_data.csv', index = False)
```

Check if there are any rows that have missing reviews

```
data = pd.read_csv('../data/cleaned_data.csv')
data.isnull().sum()
```

```
Customer Review    0
Date Of Stay       3
Customer Rating    0
Owner Responded    0
dtype: int64
```