DSBDA Assignment 9

Author Details

1. Author : Aditya Muthal
2. Roll Number : 33249
3. Batch : M10

Problem Statement

Create a review scrapper for any ecommerce website to fetch real time comments, reviews, ratings, comment tags, customer name using Python.

Notebook details

1. Number of code cells : 21
2. Python version : 3.7.4
3. Imports
    1. BeautifulSoup (bs4)
    2. pandas
    3. requests

Web page details

1. The web page used for scraping is a FlipKart product page describing headphones from Boat company.
2. The reviews, ratings, customer names, titles have been considered for scraping for the identified problem statement.

Importing required libraries

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

Setting link for scraping

```
page_link = "https://www.flipkart.com/boat-rockerz-255-pro-asap-charge-upto-40-hours-playb
```

Getting source code of target page

```
page = requests.get(page_link)
print("Status code : ", page.status_code)
```

```
Status code :  200
```

Note :

-> If the status code is not 200, then the page has not been recieved from the made request properly and further troubleshooting is required.

Initializing BeautifulSoup parser for target web page

```
soup = bs(page.content,'html.parser')
```

Scraping the web page for different fields

a) Names of customers

The names of the customers have been found to be defined inside a paragraph tag with the class named "_2sc7ZR _2V5EHH"

```
# Finding all paragraph tags with class _2sc7ZR _2V5EHH
customer_names = soup.find_all('p',class_='_2sc7ZR _2V5EHH')
```

```
customer_names
```

```
[<p class="_2sc7ZR _2V5EHH">Manash Pratim Saikai</p>,
 <p class="_2sc7ZR _2V5EHH">Himanshu  Ghotkar</p>,
 <p class="_2sc7ZR _2V5EHH">Asad Khan</p>,
 <p class="_2sc7ZR _2V5EHH">sujit das</p>,
 <p class="_2sc7ZR _2V5EHH">Pankaj Kumar</p>,
 <p class="_2sc7ZR _2V5EHH">Bhargav Rajpura</p>,
 <p class="_2sc7ZR _2V5EHH">Sagar Mallik</p>,
 <p class="_2sc7ZR _2V5EHH">Aftab Alam</p>,
 <p class="_2sc7ZR _2V5EHH">Prantik Sutradhar</p>,
 <p class="_2sc7ZR _2V5EHH">suhail k</p>]
```

```
# Extracting the customer names from the above result
customer_names_extracted = []
for i in range(0,len(customer_names)):
    customer_names_extracted.append(customer_names[i].get_text())

# printing customer names
customer_names_extracted
```

```
['Manash Pratim Saikai',
 'Himanshu  Ghotkar',
 'Asad Khan',
 'sujit das',
 'Pankaj Kumar',
 'Bhargav Rajpura',
 'Sagar Mallik',
 'Aftab Alam',
 'Prantik Sutradhar',
 'suhail k']
```

b) Ratings provided for the product

The ratings of the product have been found to be defined inside a division tag with the class named "_3LWZlK _1BLPMq"

```
product_ratings = soup.find_all('div', class_="_3LWZlK _1BLPMq")
product_ratings
```

```
[<div class="_3LWZlK _1BLPMq">5<img class="_1wB99o" src="
 <div class="_3LWZlK _1BLPMq">5<img class="_1wB99o" src="
 <div class="_3LWZlK _1BLPMq">5<img class="_1wB99o" src="
 <div class="_3LWZlK _1BLPMq">5<img class="_1wB99o" src="
 <div class="_3LWZlK _1BLPMq">4<img class="_1wB99o" src="
 <div class="_3LWZlK _1BLPMq">4<img class="_1wB99o" src="
 <div class="_3LWZlK _1BLPMq">5<img class="_1wB99o" src="
 <div class="_3LWZlK _1BLPMq">5<img class="_1wB99o" src="
 <div class="_3LWZlK _1BLPMq">4<img class="_1wB99o" src="
 <div class="_3LWZlK _1BLPMq">4<img class="_1wB99o" src="
```

```
# Extracting the product ratings from the above result
product_ratings_extracted = []
for product_rating in product_ratings:
    product_ratings_extracted.append(product_rating.get_text())

# printing product ratings
product_ratings_extracted
```

```
['5', '5', '5', '5', '4', '4', '5', '5', '4', '4']
```

c) Product reviews

The reviews of the product have been found to be defined inside a division tag with the class named "t-ZTKy"

```
product_reviews = soup.find_all('div', class_="t-ZTKy")
product_reviews
# class="t-ZTKy"
```

```
[<div class="t-ZTKy"><div><div class="">1. Material Quality - Not Bad...<br/>2. Sound
 <div class="t-ZTKy"><div><div class="">Writing this review after using more than 15
 <div class="t-ZTKy"><div><div class="">One of the best Bluetooth Boat. <br/>1- Batte
 <div class="t-ZTKy"><div><div class="">Nice product and fast delivery .  I am happy
 <div class="t-ZTKy"><div><div class="">Delivered in 20 February.. today is 28 Februa
 <div class="t-ZTKy"><div><div class="">This product is acctullly best. best coloty a
 <div class="t-ZTKy"><div><div class="">Great product., battery backup up to 14 days
 <div class="t-ZTKy"><div><div class="">Battery backup is best and full charge very q
 <div class="t-ZTKy"><div><div class="">Just got my hands on this product. Let me be
 <div class="t-ZTKy"><div><div class="">I got this product on 5th oct 2021 , today is
```

```
# Extracting product reviews from the above result
product_reviews_extracted = []
for product_review in product_reviews:
    product_reviews_extracted.append(product_review.get_text())

product_reviews_extracted
```

```
['1. Material Quality - Not Bad...2. Sound Quality - Awesome...3. Bass - Decent Quali
 "Writing this review after using more than 15 days. 255F pro+ has super sound qualit
 'One of the best Bluetooth Boat. 1- Battery Backup Amazing 2- light wait 3-Awesome S
 'Nice product and fast delivery .  I am happy nice job Flipkart thanksREAD MORE',
 "Delivered in 20 February.. today is 28 February and i didn't charge it till now sti
 'This product is acctullly best. best coloty and design.But sound is little low not
 'Great product., battery backup up to 14 days use , sound quality is great , I am ha
 'Battery backup is best and full charge very quickly. Sound quality is best .I love
 'Just got my hands on this product. Let me be honest. Build quality pretty good. Eve
 'I got this product on 5th oct 2021 , today is 15th october 2021.Iam using this proc
```

d) Review titles

The review titles of the product have been found to be defined inside a paragraph tag with the class named "_2-N8zT"

```
title = soup.find_all('p',class_='_2-N8zT')
```

```
title
```

```
[<p class="_2-N8zT">Just wow!</p>,
 <p class="_2-N8zT">Must buy!</p>,
 <p class="_2-N8zT">Great product</p>,
```

```
 <p class="_2-N8zT">Just wow!</p>,
 <p class="_2-N8zT">Good quality product</p>,
 <p class="_2-N8zT">Wonderful</p>,
 <p class="_2-N8zT">Perfect product!</p>,
 <p class="_2-N8zT">Terrific</p>,
 <p class="_2-N8zT">Good choice</p>,
 <p class="_2-N8zT">Worth the money</p>]
```

```
# Extracting review titles from above result
review_title_extracted = []
for i in range(0,len(title)):
    review_title_extracted.append(title[i].get_text())

review_title_extracted
```

```
['Just wow!',
 'Must buy!',
 'Great product',
 'Just wow!',
 'Good quality product',
 'Wonderful',
 'Perfect product!',
 'Terrific',
 'Good choice',
 'Worth the money']
```

Saving report of the above data in csv file

```
# Initializing empty dataframe
dataframe = pd.DataFrame()
```

```
dataframe
```

```
# Adding customer name to dataframe as a column
dataframe['Customer Name'] = customer_names_extracted
```

```
dataframe
```

	Customer Name
0	Manash Pratim Saikai
1	Himanshu Ghotkar
2	Asad Khan
3	sujit das

```
# Adding rest of the columns to the dataframe
dataframe['Review title'] = review_title_extracted
dataframe['Ratings'] = product_ratings_extracted
dataframe['Reviews'] = product_reviews_extracted
```

```
dataframe
```

	Customer Name	Review title	Ratings	Reviews
0	Manash Pratim Saikai	Just wow!	5	1. Material Quality - Not Bad...2. Sound Quali...
1	Himanshu Ghotkar	Must buy!	5	Writing this review after using more than 15 d...
2	Asad Khan	Great product	5	One of the best Bluetooth Boat. 1- Battery Bac...
3	sujit das	Just wow!	5	Nice product and fast delivery . I am happy n...
4	Pankaj Kumar	Good quality product	4	Delivered in 20 February.. today is 28 Februar...
5	Bhargav Rajpura	Wonderful	4	This product is acctullly best. best coloty an...
6	Sagar Mallik	Perfect product!	5	Great product., battery backup up to 14

```
# Saving the dataframe to csv
dataframe.to_csv('reviews.csv',index=True)
```

Conclusion

1. Implemented web scraper for Flipkart website product

End of Notebook