Web Scraping of Amazon Product Review

Step: 1 - Import Python Library

We will use two libraries: BeautifulSoup in bs4 and requests.

It is used commonly to scrape the web when using Python.

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

Step: 2 – Extract the HTML from the Web Page

We need to extract reviews from https://www.amazon.in/Apple-iPhone-XR-64GB-Black/product-reviews/B07JWV47JW

Save the URL in the name called 'link'

```
link = 'https://www.amazon.in/Apple-iPhone-XR-64GB-Black/product-reviews/B07JWV47JW'
```

Then we could access the content on this webpage and save the HTML in "link" by using requests.get() function in request.

```
page = requests.get(link)
print(page)
```

This would produce <Response [OK]>. Allowed to scrape.

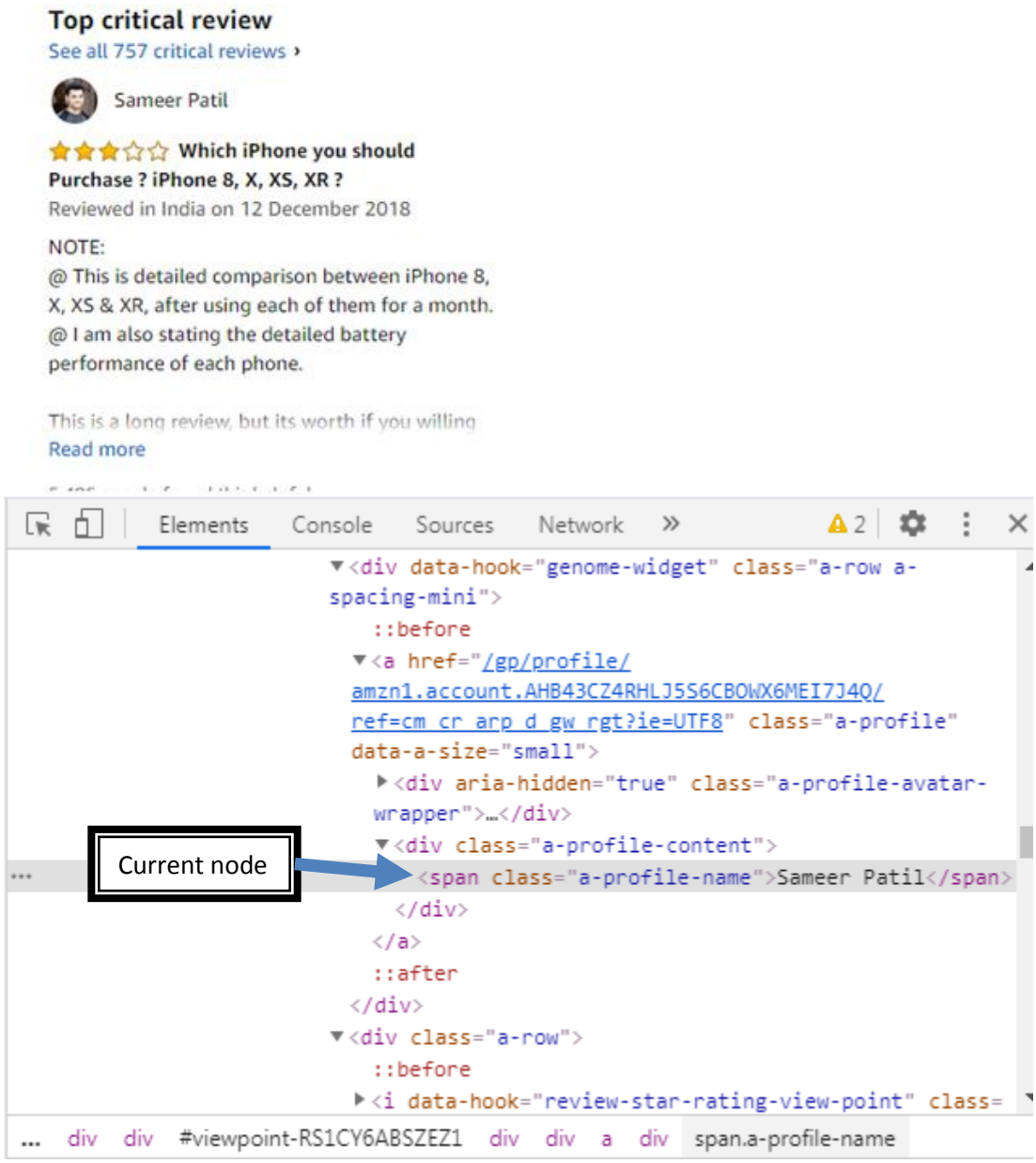
Then we apply BeautifulSoup to parse the page.

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

Now that we have the "soup", which is the raw HTML for this website, we could use a function called *prettify()* to clean the raw data and print it to see the nested structure of HTML in the "soup".

```
print(soup.prettify())
```

Step: 3 – Locate and scrape the Name, Review, Rating, Review Content.



After clicking "Inspect element" (or "Inspect", depends on different browsers), we could see the HTML of the reviewers name.

In this case, the names are located under the tag called ”span”. So we will first use the function called find_all() to find the parent node of these reviews. And then locate all elements with the tag “span” under the parent node in a loop. After finding all “span” elements, we would store them in an empty list called “name”.

```
#Names of the reviewers
name = soup.find_all('span',class_= 'a-profile-name')
name
```

```
cust_name = []
for i in range (0,len(name)):
    cust_name.append(name[i].get_text())
cust_name
```

```
['A',
 'Sameer Patil',
 'Sameer Patil',
 'Amazon Customer',
 'A',
 'Shubham Dutta',
 'Nepuni Lokho',
 'Anand S Prasad',
 'Srikanth Jatla',
 'KAILASH C.',
 'LDM',
 'Amazon Customer']
```

Remove the redundant names with pop() function,

```
cust_name.pop(0)
```

```
'Sameer Patil'
```

```
cust_name
```

```
['Sameer Patil',
 'Amazon Customer',
 'A',
 'Shubham Dutta',
 'Nepuni Lokho',
 'Anand S Prasad',
 'Srikanth Jatla',
 'KAILASH C.',
 'LDM',
 'Amazon Customer']
```

Now we get all the names from that page. Let’s see how many names have we extracted.

```
len(cust_name)
```

```
10
```

Step: 3 – Continue the same for Review Title.

```
title = soup.find_all('a', class_ = 'review-title-content')
title
```

```
[<a class="a-size-base a-link-normal review-title a-color-base review-title-content a-text-bold" data-hook="review-title" href
="/gp/customer-reviews/RS1CY6ABSZEZ1?ASIN=B07JWV47JW">
 <span>Which iPhone you should Purchase ? iPhone 8, X, XS, XR ?</span>
 </a>,
 <a class="a-size-base a-link-normal review-title a-color-base review-title-content a-text-bold" data-hook="review-title" href
="/gp/customer-reviews/R3KG909JS0ZMYX?ASIN=B07JWV47JW">
 <span>Don't buy iPhone xr from Amazon.</span>
 </a>,
 <a class="a-size-base a-link-normal review-title a-color-base review-title-content a-text-bold" data-hook="review-title" href
="/gp/customer-reviews/R3A1INE3C96MJH?ASIN=B07JWV47JW">
 <span>Happy with the purchase</span>
 </a>,
 <a class="a-size-base a-link-normal review-title a-color-base review-title-content a-text-bold" data-hook="review-title" href
="/gp/customer-reviews/R18QH2M47UUS2M?ASIN=B07JWV47JW">
 <span>Amazon is not an apple authorised reseller. Please take a note.</span>
 </a>,
 <a class="a-size-base a-link-normal review-title a-color-base review-title-content a-text-bold" data-hook="review-title" href
="/gp/customer-reviews/R2IHSOSCVXTVZ4?ASIN=B07JWV47JW">
 <span>Excellent Battery life and buttery smooth UI</span>
 </a>,
 <a class="a-size-base a-link-normal review-title a-color-base review-title-content a-text-bold" data-hook="review-title" href
="/gp/customer-reviews/RJFHMJ147WZP6?ASIN=B07JWV47JW">
 <span>Never purchase a phone online.</span>
```

```
review_title = []
for i in range (0,len(title)):
    review_title.append(title[i].get_text())
review_title
```

```
['\nWhich iPhone you should Purchase ? iPhone 8, X, XS, XR ?\n',
 "\nDon't buy iPhone xr from Amazon.\n",
 '\nHappy with the purchase\n',
 '\nAmazon is not an apple authorised reseller. Please take a note.\n',
 '\nExcellent Battery life and buttery smooth UI\n',
 '\nNever purchase a phone online.\n',
 '\nAwesome iPhone\n',
 '\nExchange of I phone xr with xs.\n',
 "\nWorth upgrading from the 6S? I'll know in a few months\n",
 '\nDef a bad experience\n']
```

Now, remove '\n' on both the sides,

```
review_title[:] = [titles.lstrip('\n') for titles in review_title]
review_title[:]
```

```
review_title[:] = [titles.rstrip('\n') for titles in review_title]
review_title[:]
```

```
['Which iPhone you should Purchase ? iPhone 8, X, XS, XR ?',
 "Don't buy iPhone xr from Amazon.",
 'Happy with the purchase',
 'Amazon is not an apple authorised reseller. Please take a note.',
 'Excellent Battery life and buttery smooth UI',
 'Never purchase a phone online.',
 'Awesome iPhone',
 'Exchange of I phone xr with xs.',
 "Worth upgrading from the 6S? I'll know in a few months",
 'Def a bad experience']
```

Step: 4 - Now that we have our required data in the form of a list, we will be using Python **Pandas** library to save the data in an Excel file. Before that, we have to convert the list into a DataFrame.

```
: import pandas as pd
```

```
: df = pd.DataFrame()
  df['Customer Name'] = cust_name
  df['Review Title'] = review_title
  df
```

:

	Customer Name	Review Title
0	Sameer Patil	Which iPhone you should Purchase ? iPhone 8, X...
1	Amazon Customer	Don't buy iPhone xr from Amazon.
2	A	Happy with the purchase
3	Shubham Dutta	Amazon is not an apple authorised reseller. Pl...
4	Nepuni Lokho	Excellent Battery life and buttery smooth UI
5	Anand S Prasad	Never purchase a phone online.
6	Srikanth Jatla	Awesome iPhone
7	KAILASH C.	Exchange of I phone xr with xs.
8	LDM	Worth upgrading from the 6S? I'll know in a fe...
9	Amazon Customer	Def a bad experience

Source: https://youtu.be/ecAJfHHppVs