Sinhgad Institutes

Name of the Student:_____**Roll no:**_____

CLASS:- T.E.[I.T] **Division: A**

Course: - 2019

Subject: 314457: Data Science and Big Data Analytics Laboratory

PART_ C _Assignment No. 01

Marks: _____/10

Date of Performance: ___/____/_____ **Sign with Date: _____**

Part- C

ASSIGNMENT NO: 01

TITLE:

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

AIM:

To create a review scrapper for ecommerce website.

OBJECTIVE:

1. To scrap an ecommerce website.
2. https://www.thewhiskyexchange.com/c/35/japanese-whisky

We are going to scrape this website to retrieve the information from there.

Software used: Python Shell 3.7.3

THEORY:

Scraping product reviews from ecommerce websites has become one of the most vital competitive intelligence activities in and around the ecommerce space. Product reviews on ecommerce websites are a great source of unbiased reviews organically left by actual consumers. This means, if you are a manufacturer trying to gather deep insights about your product, look no further than ecommerce product review pages. While the availability of product reviews on various eCommerce sites is vast and deep, not every company has the expertise, infrastructure, and resources to crawl and extract reviews from eCommerce sites in an automated manner.

Luckily, PromptCloud specializes in large-scale web scraping solutions and can help you with scraping product reviews from eCommerce portals of your choice. Our fully managed service comes with end-to-end crawler management which would insulate you from the nuances of web data extraction activity and help you focus on the application of the delivered data.

A very simple pie-chart is created using just the input vector and labels. The below script will create and save the pie chart in the current R working directory.

Applications of eCommerce product reviews scraping: eCommerce is taking over the world by storm and there's no scarcity of product reviews on these eCommerce portals. The main advantage to manufacturing companies is the unbiased nature of these reviews which will help them understand their consumers so as to serve them better. Here are the most popular applications of product reviews scraping.

Understand your consumer preferences: Staying up-to-date with customer preferences is crucial to successful products. If your product isn't addressing what the consumer wants, you are leaving money on the table and contributing to customer dissatisfaction. To avoid this, understanding the demands and needs of your target customer is vital. This is where product reviews scraping can fill in by helping you listen to the customer's voice. By extracting product reviews and analyzing it with the right goals, your business can understand the crucial factors that are driving sales in your niche and tweak the products accordingly.

Brand monitoring: Reputation or brand image is a huge factor when it comes to customer loyalty and business growth for any organization. Maintaining a positive brand image is crucial and this means your business should be all ears to the customer grievances. Brand monitoring can help you detect unaddressed issues of your customers that can escalate to bigger issues if not detected early on. Scraping reviews from eCommerce sites can help you monitor this and preserve your positive brand image.

Competitor analysis: Listening to your customers alone is just not enough to wade through this ever so competitive business world of today. Sometimes, data associated to your competitors can help you detect low hanging fruits which you can leverage before they do. For example, if reviews on your competitors products indicate the demand for a particular feature, be the first one to incorporate that to your own product. This would help you stay ahead of the curve and garner more users.

Natural language processing: Natural language processing (NLP) is all about enabling machines to understand the context behind human languages. Systems developed using NLP helps run voice assistant platforms like Siri, Google Now and Cortana, translation services and artificial intelligence systems. Huge amounts of data is a must to cater to the requirements of a NLP system. Since product reviews scraped from eCommerce sites is user-generated content, it makes for the perfect data for training Natural language processing systems.

Fraud Detection: Counterfeit products have always been a threat to brands. Not only do they affect sales figures, but also leave a bad impression among customers who may never realize the faulty product they received was in fact a fake.

Scraping product reviews helps you access this data which might have hints about some ongoing fraud. An alarming number of negative reviews are definitely worth investigating further to rule out counterfeit products being sold by the supplier. Near real-time crawls or live crawls can be performed to identify the ecommerce partners who don't stick to the agreement.

Setup the Scraping Project

Our setup is pretty simple. Just create a folder and install Beautiful Soup, pandas, and requests. To create a folder and install the libraries, enter the commands given below.

```
mkdir scraper
pip install beautifulsoup4
pip install requests
pip install pandas
```

Now, create a file inside that folder and name it anything you like. I am using the name scraper.py. We are going to import requests, pandas, and bs4.

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

Now, we are going to set the base URL of the main page because we'll need that when we construct our URLs for each of the individual products.

Also, we will send a user-agent on every HTTP request, because if you make GET request using requests then by default the user-agent is Python which might get blocked.

So, to override that, we will declare a variable which will store our user-agent.

```
baseurl = "https://www.thewhiskyexchange.com"
headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/89.0.4389.82 Safari/537.36'}
```

Now we need to investigate the page so that we can figure out where the links are and how we're going to get them.

CONCLUSION: After the study of this assignment we have learnt Scrape the website by using python libraries.

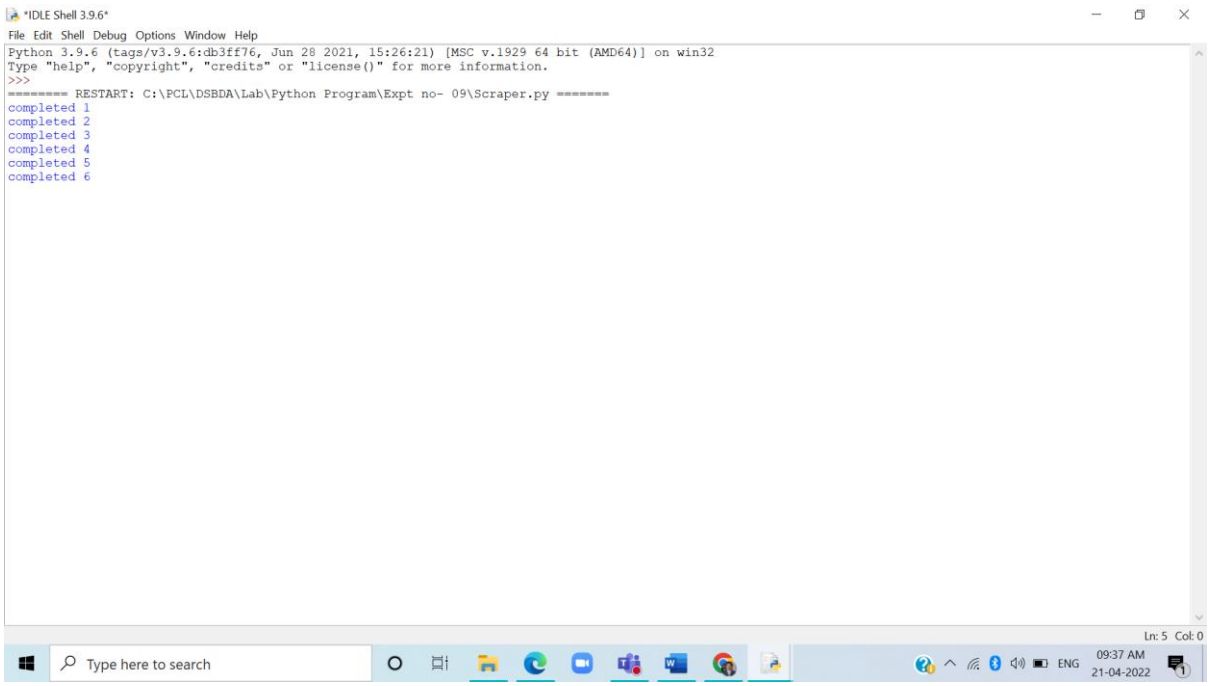
Python Program

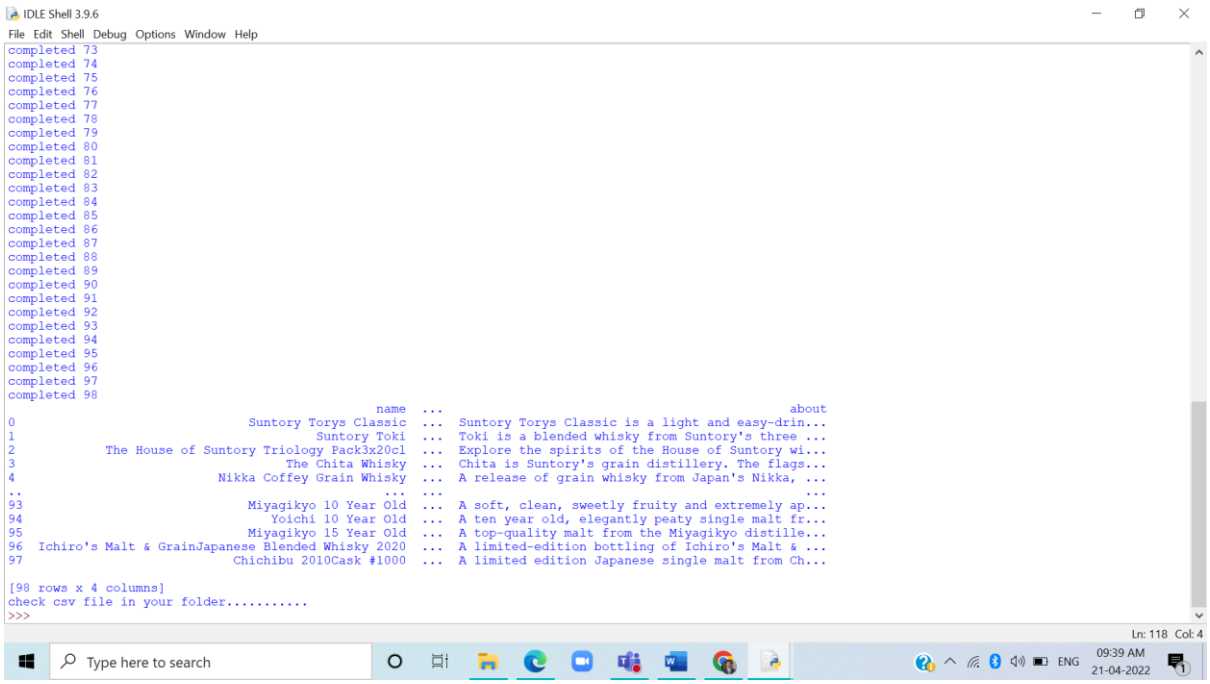
```
##https://www.freecodecamp.org/news/scraping-ecommerce-
website-with-python/
##https://rentechdigital.com/smartscraper/review-scraping

import requests
from bs4 import BeautifulSoup
import pandas as pd

baseurl = "https://www.thewhiskyexchange.com"
headers = {'User-Agent': 'Mozilla/5.0 (Windows NT 10.0; Win64;
x64)      AppleWebKit/537.36      (KHTML,      like      Gecko)
Chrome/89.0.4389.82 Safari/537.36'}
productlinks = []
t={}
data=[]
c=0
for x in range(1,6):
     k=requests.get('https://www.thewhiskyexchange.com/c/35/ja
     panese-whisky?pg={}&psize=24&sort=pasc'.format(x)).text
     soup=BeautifulSoup(k,'html.parser')
     productlist=soup.find_all("li",{"class":"product-
grid__item"})
     print(productlist)

for product in productlist:
   link=product.find("a",{"class":"product-card"}).get('href')
        productlinks.append(baseurl + link)

for link in productlinks:
    f = requests.get(link,headers=headers).text
    hun=BeautifulSoup(f,'html.parser')
    try:
        price=hun.find("p",{"class":"product-
```

```
action__price"}).text.replace('\n',"")
    except:
        price = None
    try:
        about=hun.find("div",{"class":"product-
main__description"}).text.replace('\n',"")
    except:
        about=None
    try:
        rating=hun.find("div",{"class":"review-
overview"}).text.replace('\n',"")
    except:
        rating=None
    try:
        name=hun.find("h1",{"class":"product-
main__name"}).text.replace('\n',"")
    except:
        name=None
    whisky=
{"name":name,"price":price,"rating":rating,"about":about}
    data.append(whisky)
    c=c+1
    print("completed",c)
df = pd.DataFrame(data)
print(df)
df.to_csv('review scraper.csv')
print("check csv file in your folder..........")
```

>>>


IDLE Shell 3.9.6
File Edit Shell Debug Options Window Help
Python 3.9.6 (tags/v3.9.6:db3ff76, Jun 28 2021, 15:26:21) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
======== RESTART: C:\PCL\DSBDA\Lab\Python Program\Expt no- 09\Scraper.py ========
Ln: 123 Col: 4
Type here to search
ENG
11:39 AM
21-04-2022


IDLE Shell 3.9.6
File Edit Shell Debug Options Window Help
Python 3.9.6 (tags/v3.9.6:db3ff76, Jun 28 2021, 15:26:21) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license()" for more information.
>>>
======== RESTART: C:\PCL\DSBDA\Lab\Python Program\Expt no- 09\Scraper.py ========
completed 1
completed 2
completed 3
completed 4
completed 5
completed 6
Ln: 5 Col: 0
Type here to search
ENG
09:37 AM
21-04-2022


IDLE Shell 3.9.6
File Edit Shell Debug Options Window Help
completed 73
completed 74
completed 75
completed 76
completed 77
completed 78
completed 79
completed 80
completed 81
completed 82
completed 83
completed 84
completed 85
completed 86
completed 87
completed 88
completed 89
completed 90
completed 91
completed 92
completed 93
completed 94
completed 95
completed 96
completed 97
completed 98
name ... about
0 Suntory Torys Classic ... Suntory Torys Classic is a light and easy-drin...
1 Suntory Toki ... Toki is a blended whisky from Suntory's three ...
2 The House of Suntory Triology Pack3x20cl ... Explore the spirits of the House of Suntory wi...
3 The Chita Whisky ... Chita is Suntory's grain distillery. The flags...
4 Nikka Coffey Grain Whisky ... A release of grain whisky from Japan's Nikka, ...
.. ..
93 Miyagikyo 10 Year Old ... A soft, clean, sweetly fruity and extremely ap...
94 Yoichi 10 Year Old ... A ten year old, elegantly peaty single malt fr...
95 Miyagikyo 15 Year Old ... A top-quality malt from the Miyagikyo distille...
96 Ichiro's Malt & GrainJapanese Blended Whisky 2020 ... A limited-edition bottling of Ichiro's Malt & ...
97 Chichibu 2010Cask #1000 ... A limited edition Japanese single malt from Ch...
[98 rows x 4 columns]
check csv file in your folder.....
>>>
Ln: 118 Col: 4
Type here to search
ENG
09:39 AM
21-04-2022