PYTHON PROJECT REPORT

BY

BATCH 2019 – 2020

INDHU ARIVALAGAN
PRASHANTH RATHINAVEL
YOGITHASATYASAI PANTHAM

ABSTRACT

Web scraping services are provided by computer software which extracts the required facts from the website. Web scraping services mainly aim at converting unstructured data collected from the websites into structured data which can be stockpiled and scrutinized in a centralized databank. Therefore, web scraping services have a direct influence on the outcome of the reason as to why the data collected in necessary.

IMPORTANCE OF WEB SCRAPING

Web scraping services have gone a long way in the provision of very useful information to various organizations. But business companies are the ones that benefit more from web scraping services. Some of the benefits associated with web scraping services are:

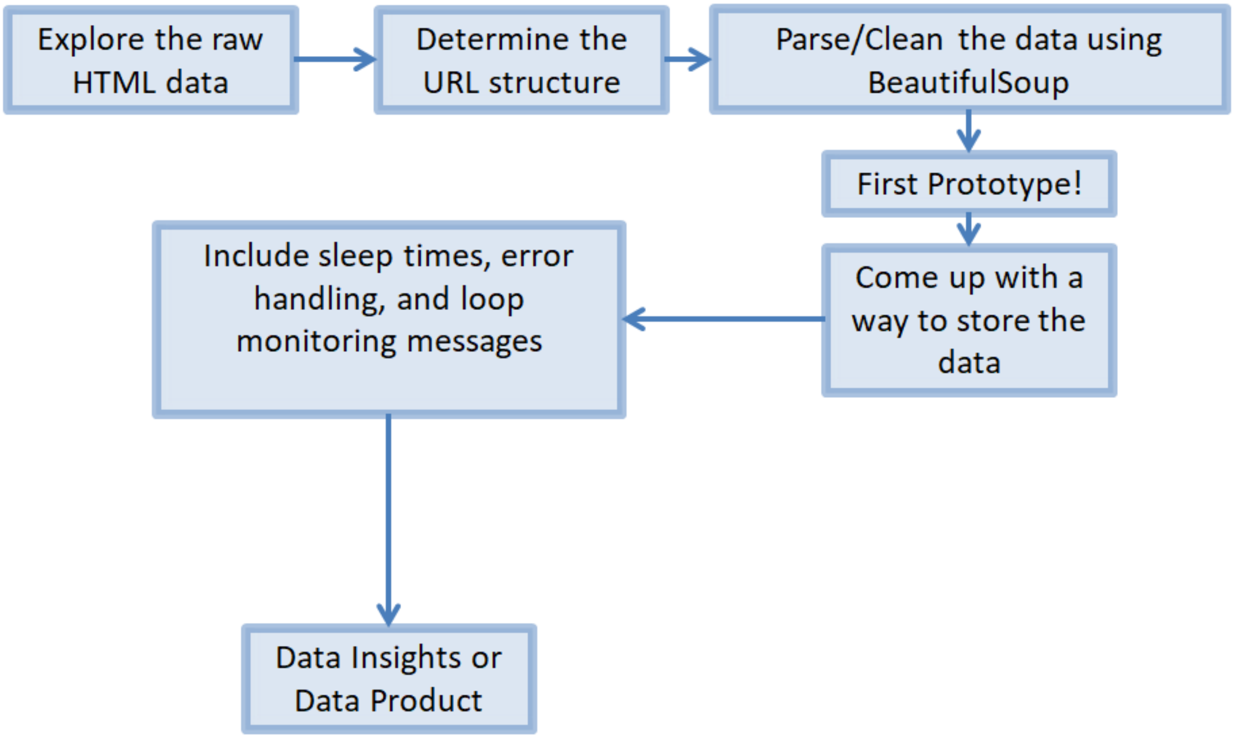
> Helps the firms to easily send notifications to their customers including price changes, promotions, an introduction of a new product into the market. Etc.

> It enables firms to compare their product prices with those of their competitors

> It helps the meteorologists to monitor weather changes thus being able to focus weather conditions more efficiently

> It also assists researchers with extensive information about peoples’ habits among many others.

FLOW CHART


Explore the raw HTML data
Determine the URL structure
Parse/Clean the data using BeautifulSoup
First Prototype!
Come up with a way to store the data
Include sleep times, error handling, and loop monitoring messages
Data Insights or Data Product

STEP 1: Analyzing the scrapping page and its objects

We are extracting the details of the python books in Amazon website (.fr region) by generating URLs using ASIN and ISBN. Below are few URLs whose information were scrapped,

STEP 2: Extracting data using beautiful soup library and using json libraries to convert it into JSON files.
The data from the website is HTML and mostly unstructured. Hence, in this step,

- The web scraper will parse and extract structured data from the downloaded contents.
- Our project extracts the data and save that into JSON and then build a database to read JSON and displays them in the console.

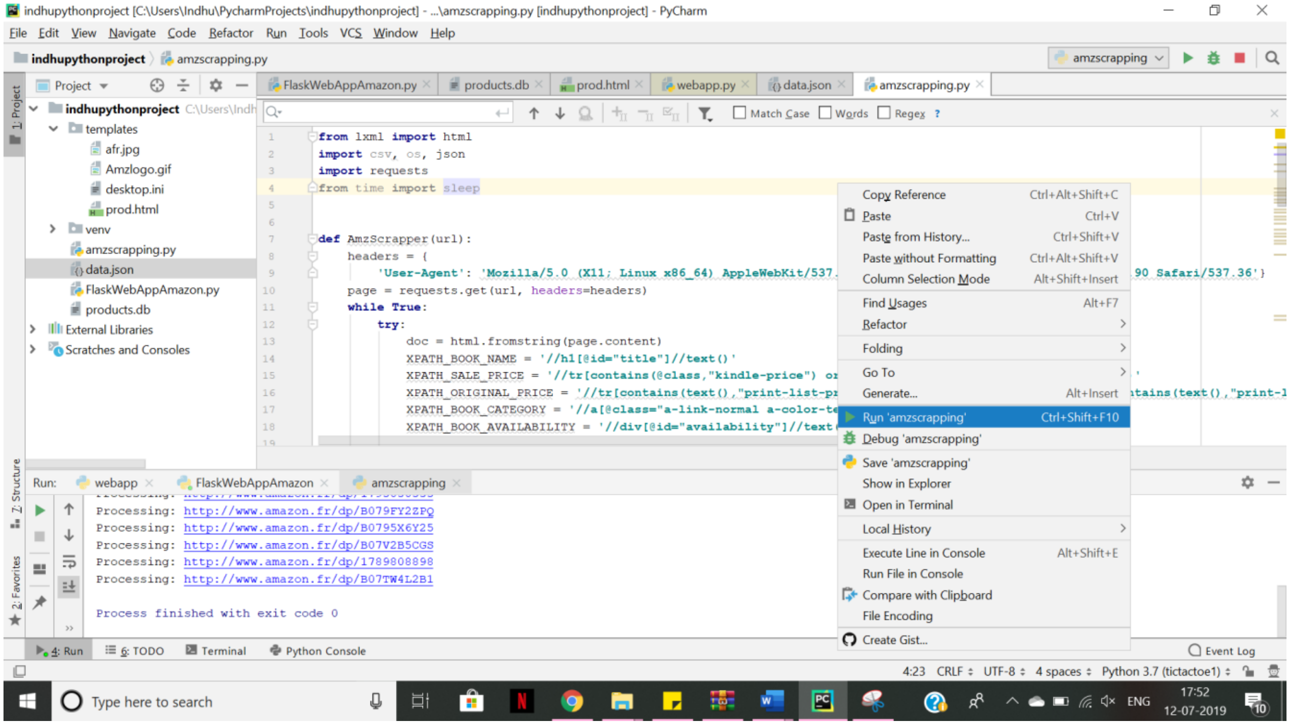
The following data points will be displayed in the JSON file,

> Book Name
> Sale Price
> Original Price
> Book Category
> Book Availability

STEP 3: After successful completion of scrapping the generated web pages, the output with above mentioned data points will be displayed in the console.

EXECUTION OF THE WEB SCRAPPING

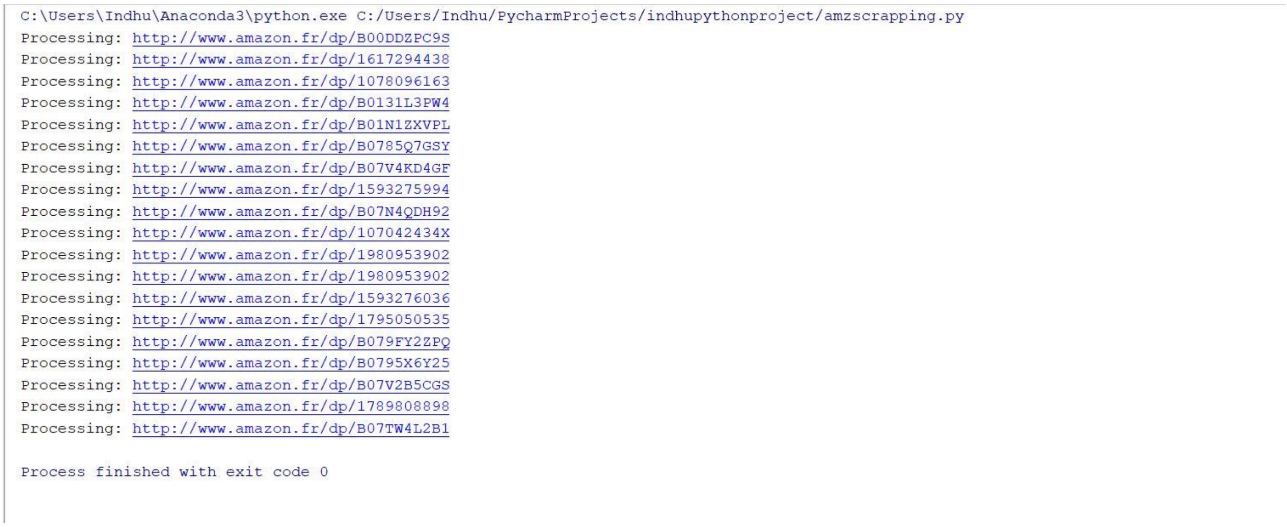
Step 1: Right click on the amzscrapping.py and run the application



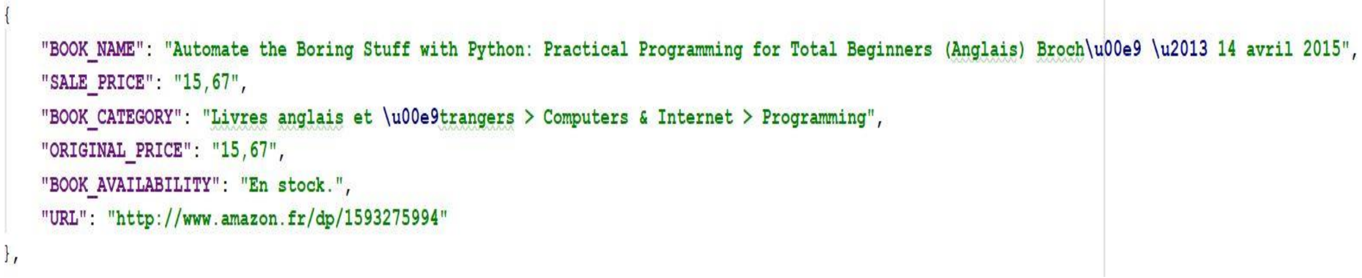
```
from lxml import html
import csv, os, json
import requests
from time import sleep


def AmzScrapper(url):
    headers = {
        'User-Agent': 'Mozilla/5.0 (X11; Linux x86_64) AppleWebKit/537.                              90 Safari/537.36'}
    page = requests.get(url, headers=headers)
    while True:
        try:
            doc = html.fromstring(page.content)
            XPATH_BOOK_NAME = '//h1[@id="title"]//text()'
            XPATH_SALE_PRICE = '//tr[contains(@class,"kindle-price") or
            XPATH_ORIGINAL_PRICE = '//tr[contains(text(),"print-list-pr                              tains(text(),"print-l
            XPATH_BOOK_CATEGORY = '//a[@class="a-link-normal a-color-te
            XPATH_BOOK_AVAILABILITY = '//div[@id="availability"]//text
```

```
Processing: http://www.amazon.fr/dp/B079FY2ZPQ
Processing: http://www.amazon.fr/dp/B0795X6Y25
Processing: http://www.amazon.fr/dp/B07V2B5CGS
Processing: http://www.amazon.fr/dp/1789808898
Processing: http://www.amazon.fr/dp/B07TW4L2B1

Process finished with exit code 0
```

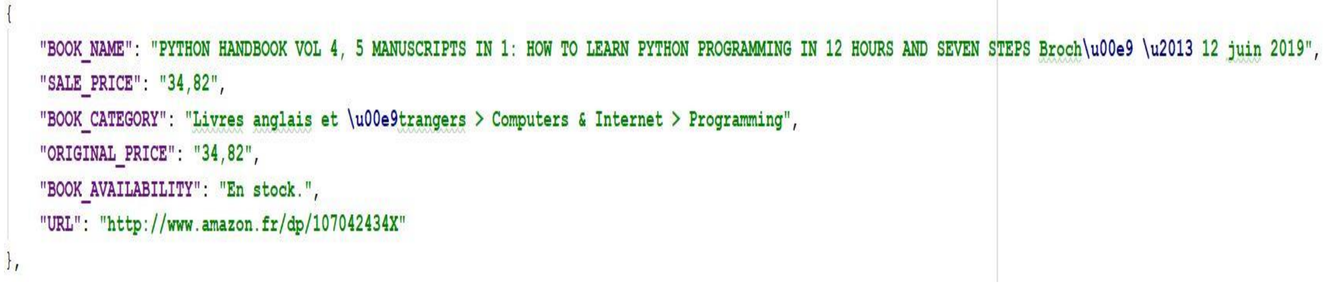
Step 2: click on the “data.json” file that is generated to scrape the data of the Python Books.



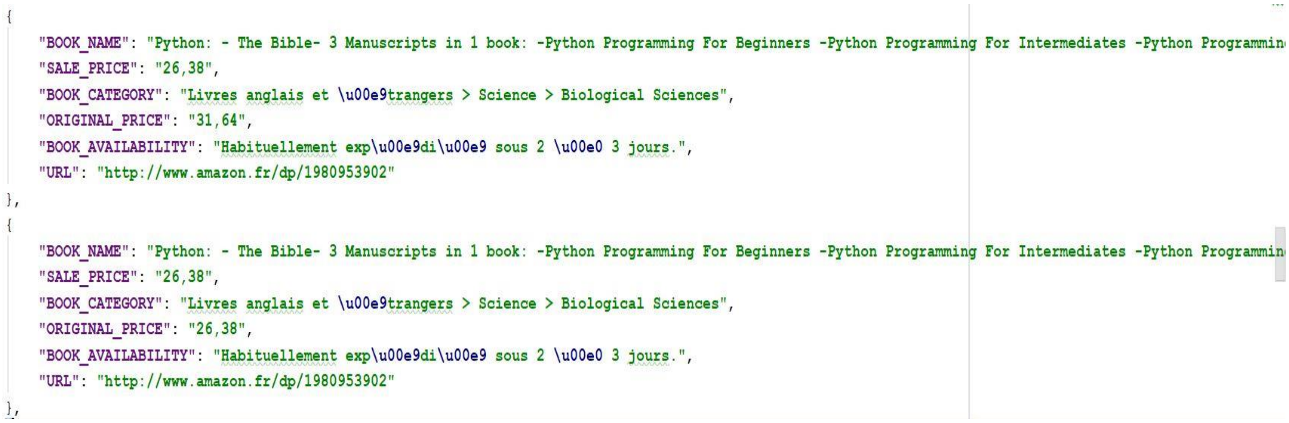
```
C:\Users\Indhu\Anaconda3\python.exe C:/Users/Indhu/PycharmProjects/indhupythonproject/amzscrapping.py
Processing: http://www.amazon.fr/dp/B00DDZPC9S
Processing: http://www.amazon.fr/dp/1617294438
Processing: http://www.amazon.fr/dp/1078096163
Processing: http://www.amazon.fr/dp/B0131L3PW4
Processing: http://www.amazon.fr/dp/B01N1ZXVPL
Processing: http://www.amazon.fr/dp/B0785Q7GSY
Processing: http://www.amazon.fr/dp/B07V4KD4GF
Processing: http://www.amazon.fr/dp/1593275994
Processing: http://www.amazon.fr/dp/B07N4QDH92
Processing: http://www.amazon.fr/dp/107042434X
Processing: http://www.amazon.fr/dp/1980953902
Processing: http://www.amazon.fr/dp/1980953902
Processing: http://www.amazon.fr/dp/1593276036
Processing: http://www.amazon.fr/dp/1795050535
Processing: http://www.amazon.fr/dp/B079FY2ZPQ
Processing: http://www.amazon.fr/dp/B0795X6Y25
Processing: http://www.amazon.fr/dp/B07V2B5CGS
Processing: http://www.amazon.fr/dp/1789808898
Processing: http://www.amazon.fr/dp/B07TW4L2B1

Process finished with exit code 0
```



```
{
    "BOOK_NAME": "Automate the Boring Stuff with Python: Practical Programming for Total Beginners (Anglais) Broch\u00e9 \u2013 14 avril 2015",
    "SALE_PRICE": "15,67",
    "BOOK_CATEGORY": "Livres anglais et \u00e9trangers > Computers & Internet > Programming",
    "ORIGINAL_PRICE": "15,67",
    "BOOK_AVAILABILITY": "En stock.",
    "URL": "http://www.amazon.fr/dp/1593275994"
},
```



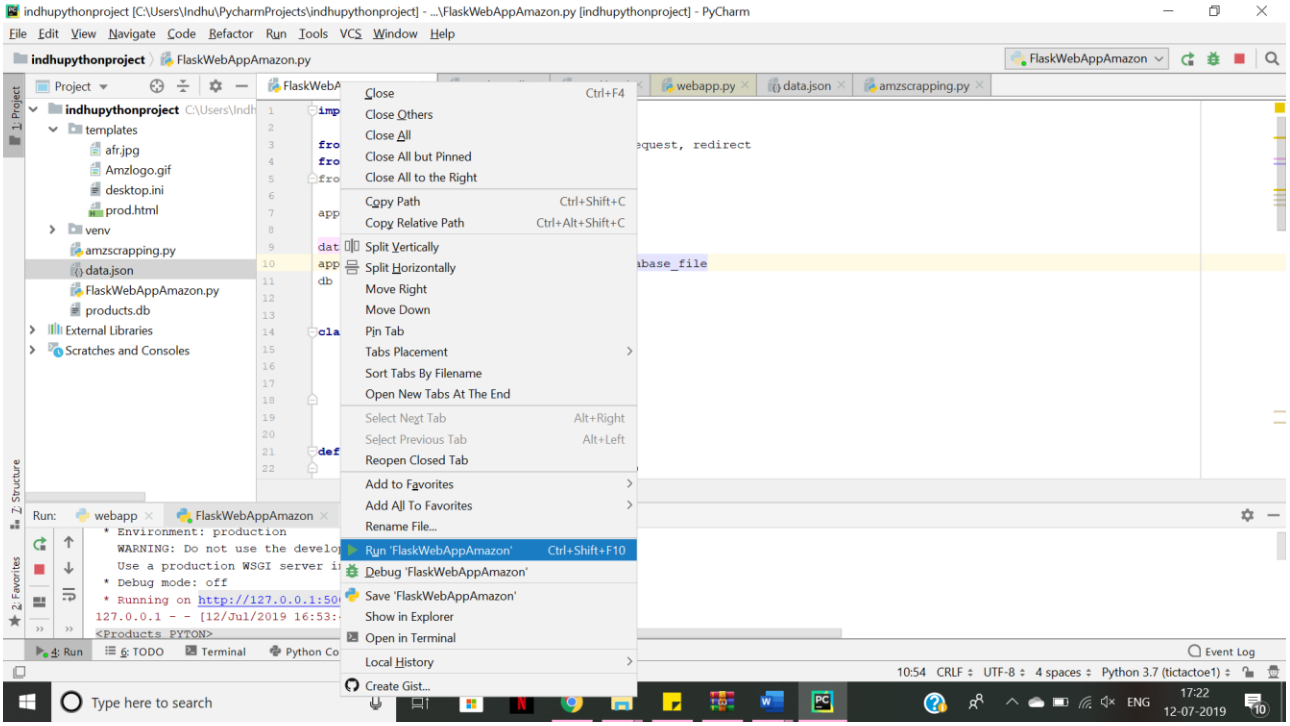
```
{
    "BOOK_NAME": "PYTHON HANDBOOK VOL 4, 5 MANUSCRIPTS IN 1: HOW TO LEARN PYTHON PROGRAMMING IN 12 HOURS AND SEVEN STEPS Broch\u00e9 \u2013 12 juin 2019",
    "SALE_PRICE": "34,82",
    "BOOK_CATEGORY": "Livres anglais et \u00e9trangers > Computers & Internet > Programming",
    "ORIGINAL_PRICE": "34,82",
    "BOOK_AVAILABILITY": "En stock.",
    "URL": "http://www.amazon.fr/dp/107042434X"
},
```



```
{
    "BOOK_NAME": "Python: - The Bible- 3 Manuscripts in 1 book: -Python Programming For Beginners -Python Programming For Intermediates -Python Programmin
    "SALE_PRICE": "26,38",
    "BOOK_CATEGORY": "Livres anglais et \u00e9trangers > Science > Biological Sciences",
    "ORIGINAL_PRICE": "31,64",
    "BOOK_AVAILABILITY": "Habituellement exp\u00e9di\u00e9 sous 2 \u00e0 3 jours.",
    "URL": "http://www.amazon.fr/dp/1980953902"
},
{
    "BOOK_NAME": "Python: - The Bible- 3 Manuscripts in 1 book: -Python Programming For Beginners -Python Programming For Intermediates -Python Programmin
    "SALE_PRICE": "26,38",
    "BOOK_CATEGORY": "Livres anglais et \u00e9trangers > Science > Biological Sciences",
    "ORIGINAL_PRICE": "26,38",
    "BOOK_AVAILABILITY": "Habituellement exp\u00e9di\u00e9 sous 2 \u00e0 3 jours.",
    "URL": "http://www.amazon.fr/dp/1980953902"
},
```

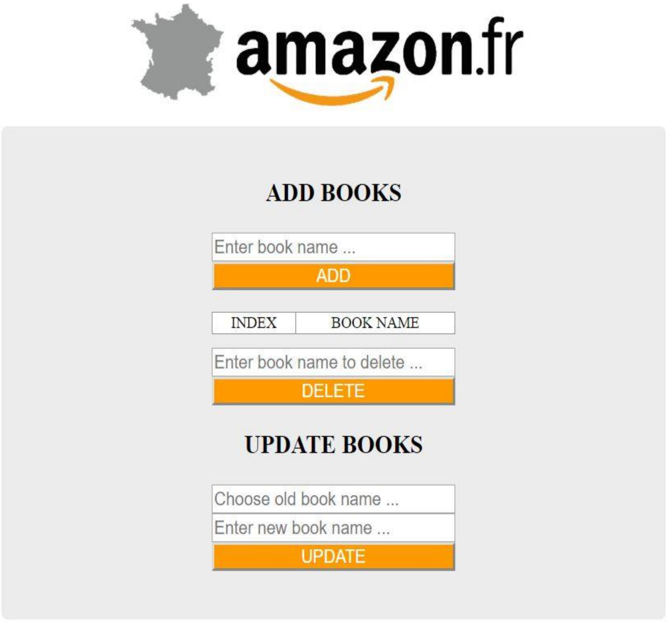
EXECUTION OF THE WEB APPLICATION

Step 1: Right click on the FlaskWebappAmazon and run the application

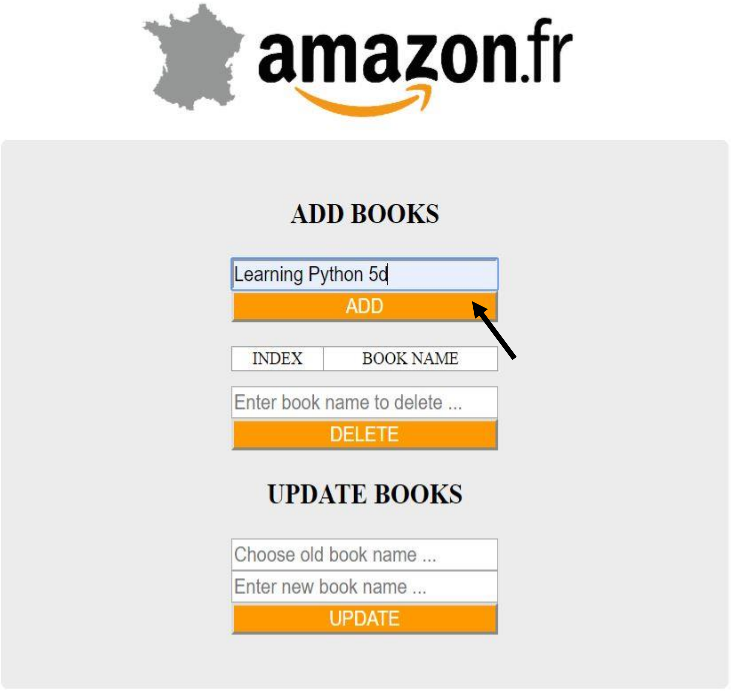


```
 * Environment: production
   WARNING: Do not use the develo
   Use a production WSGI server i
 * Debug mode: off
 * Running on http://127.0.0.1:50
127.0.0.1 - - [12/Jul/2019 16:53:
<Products PYTON>
```

Step 2: Welcome Screen looks like this

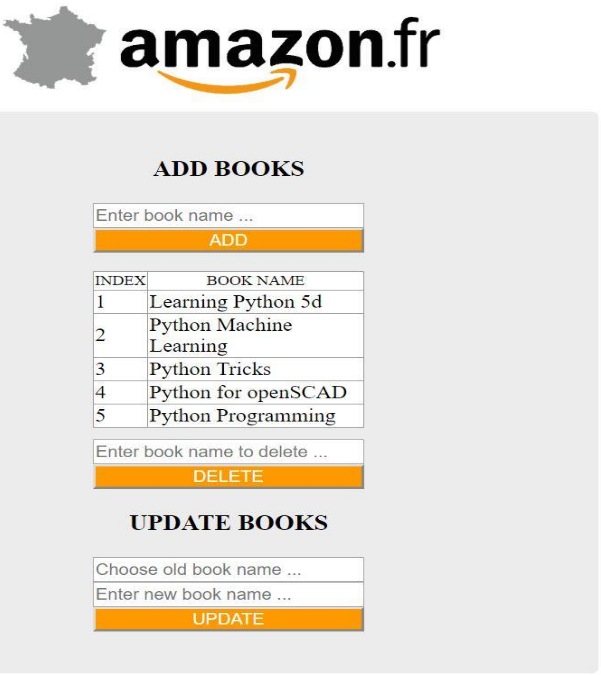


Step 3: Enter the name of the Python Book you would like to add



Learning Python 5d

Step 4: Please find below the Added Python Book List.

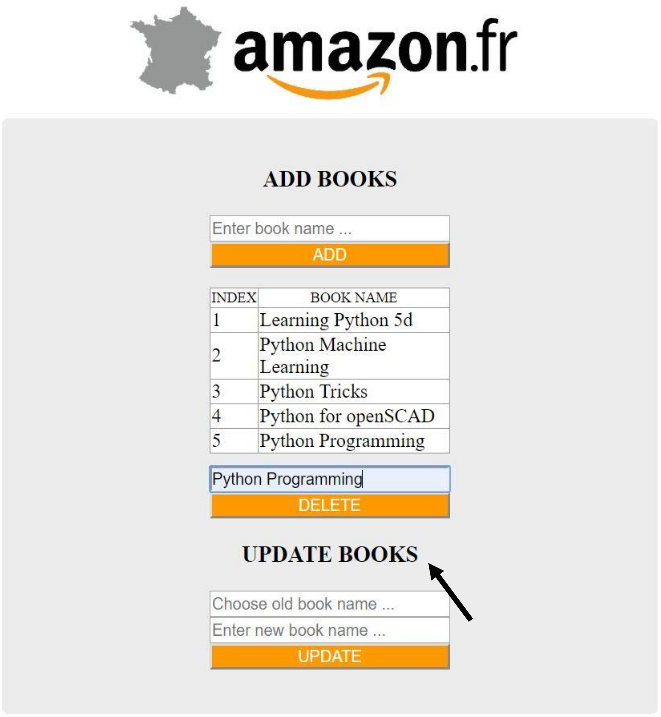


ADD BOOKS

INDEX	BOOK NAME
1	Learning Python 5d
2	Python Machine Learning
3	Python Tricks
4	Python for openSCAD
5	Python Programming

UPDATE BOOKS

Step 5: please enter the name of the python book you want to delete



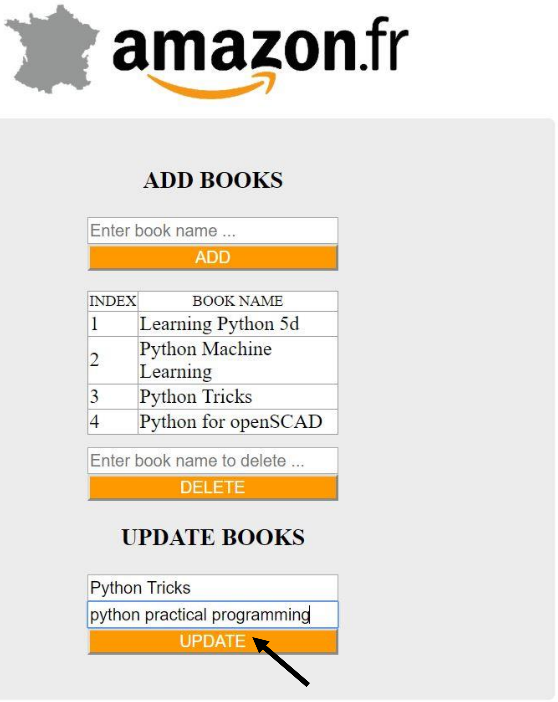
ADD BOOKS

INDEX	BOOK NAME
1	Learning Python 5d
2	Python Machine Learning
3	Python Tricks
4	Python for openSCAD
5	Python Programming

Python Programming

UPDATE BOOKS

Step 6: Python Programming book is deleted and please enter the name of the python book you want to update now

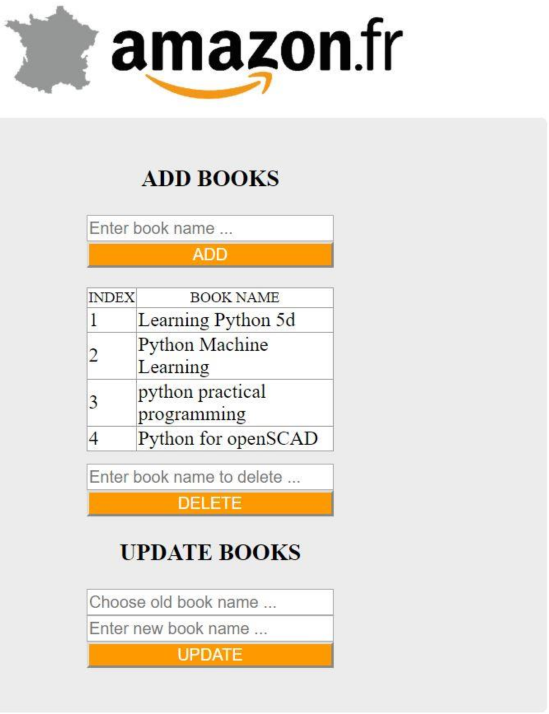


INDEX	BOOK NAME
1	Learning Python 5d
2	Python Machine Learning
3	Python Tricks
4	Python for openSCAD

Python Tricks

python practical programming

Step 7: The final page of the output shows the following



INDEX	BOOK NAME
1	Learning Python 5d
2	Python Machine Learning
3	python practical programming
4	Python for openSCAD

ADVANTAGES OF WEB SCRAPING

The following are some of the advantages of using web scraping services

- Automation of the data
- Web scraping can retrieve both static and dynamic web pages
- Page contents of various websites can be transformed
- It allows formulation of vertical aggregation platforms thus even complicated data can still be extracted from different websites.
- Web scraping programs recognize semantic annotation
- All the required data can be retrieved from their websites
- The data collected is accurate and reliable

LIMITATIONS OF WEB SCRAPING

- High volume of web scraping can causer regulatory damage to the pages
- Scale of measure; the scales of the web scraper can differ with the units of measure of the source file thus making it somewhat hard for the interpretation of the data
- The level of source complexity; if the information being extracted is very complicated, web scraping will also be paralyzed.

Besides web scraping providing useful data and information, it experiences several challenges. The good thing is that the web scraping services providers are always improvising techniques to ensure that the information gathered is accurate, timely, reliable and treated with the highest levels of confidentiality.