Web Mining Lab Assignment 2

Name: Om Ashish Mishra

Registration Number: 16BCE0789

Slot: F2

# The Question:

Use scrapy or Beautiful Soup to crawl any one of the E-commerce website of your choice and perform the same. The following information needs to be extracted from the page:

(Choose any one product : e.g laptop, smartphone … etc)

• Product Name • Product price • Product discount • Product image

# The Sample Answers:

## Sample Try Given By Sir:

**The Code:**

from html.parser import HTMLParser

from urllib.request import urlopen

from urllib import parse

class LinkParser(HTMLParser):

def handle\_starttag(self, tag, arrrs):

if tag == 'a':

for (key,value) in attrs:

if key == 'href':

newUrl = parse.urljoin(self.baseUrl, value)

self.links = self.links + [newUrl]

def getLinks(self, url):

self.links = []

self.baseUrl = url

response = urlopen(url)

if response.getheader('Content-type')=='text/html':

htmlBytes = response.read()

htmlString = htmlBytes.decode("utf-8")

self.feed(htmlString)

return htmlString, self.links

else:

return "",[]

def spider(url, word, maxPages):

pagesToVisit = [url]

numberVisited = 0

foundWord = False

while numberVisited < maxPages and pagesToVisit != [] and not foundWord:

numberVisited = numberVisited + 1

url = pagesToVisit[0]

pagesToVisit = pagesToVisit[1:]

try:

print(numberVisited, "Visiting:", url)

parser = LinkParser()

data, links = parser.getLinks(url)

if data.find(work) > -1:

foundWound = True

pagesToVisit = pagesToVisit + links

print(" \*\*Sucess!\*\* ")

expect:

print(" \*\*Failed!\*\* ")

if foundWord:

print("The word", word, "was found at", url)

else:

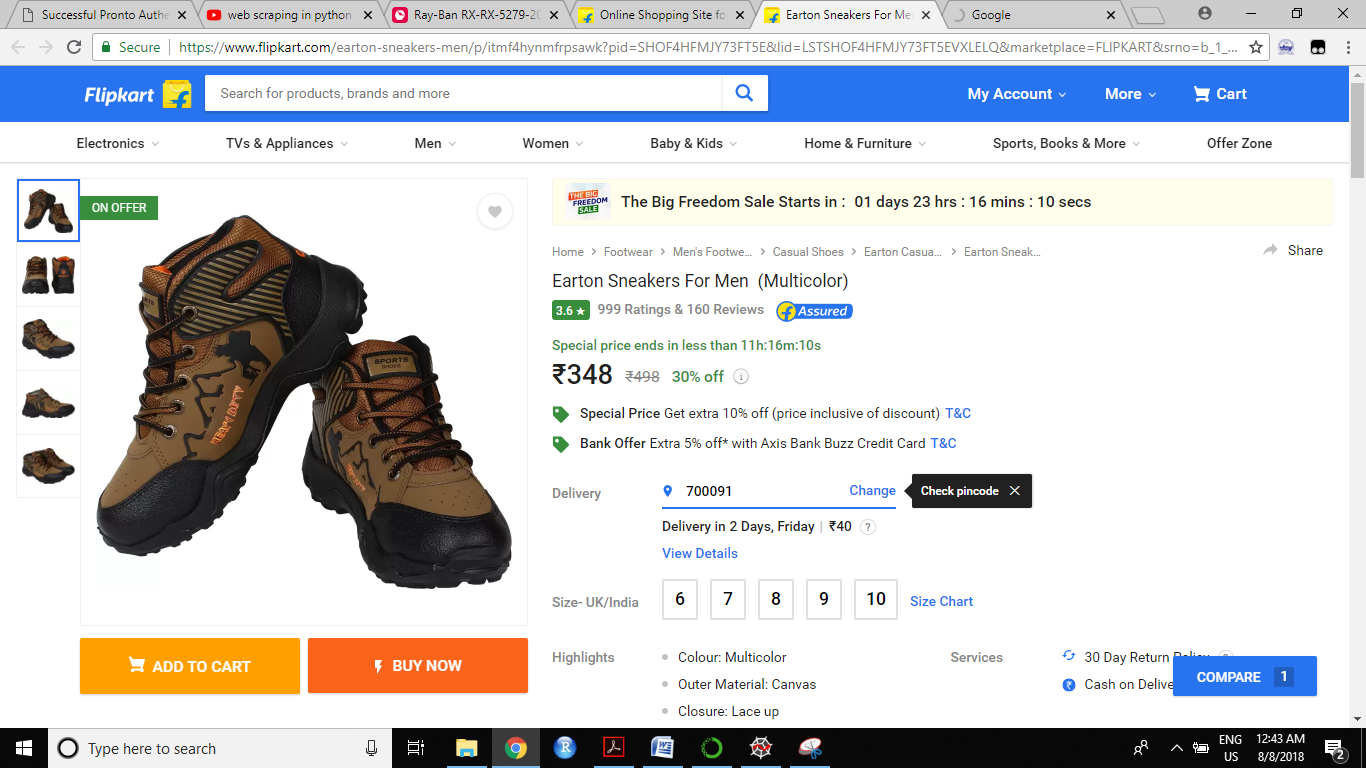
print("Word never found")

The Output:

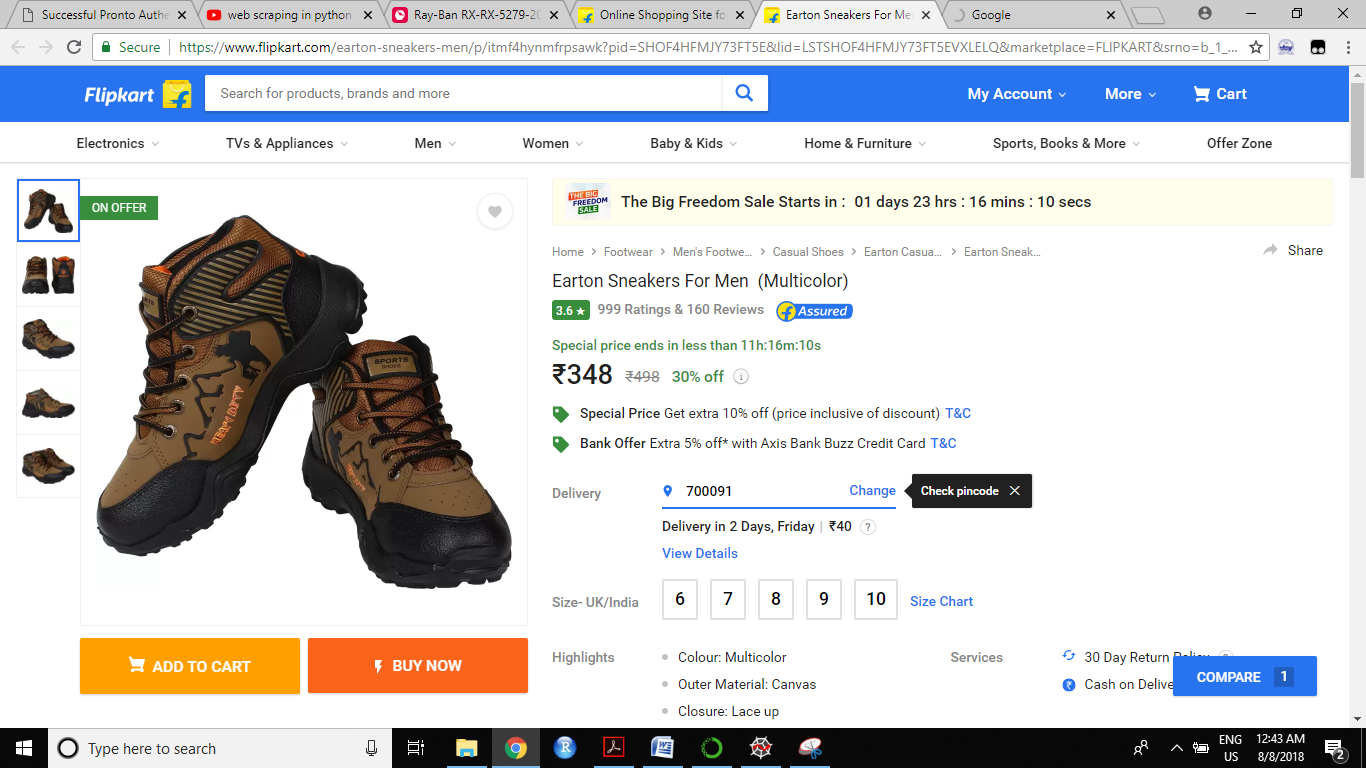


**The Actual Answer:**

**The Page I used:**

****

**The Representation of Various Entities:**

****

1. PRODUCT NAME

The Code:

from nltk.tokenize import word\_tokenize

from nltk.corpus import stopwords

from bs4 import BeautifulSoup

import urllib.request

stop\_words = stopwords.words('english')

response = urllib.request.urlopen('https://www.flipkart.com/earton-sneakers-men/p/itmf4hynmfrpsawk?pid=SHOF4HFMJY73FT5E&lid=LSTSHOF4HFMJY73FT5EVXLELQ&marketplace=FLIPKART&srno=b\_1\_1&otracker=hp\_omu\_Deals+of+the+Day\_4\_Under+%E2%82%B9699%2BExtra+10%25\_IB7FUTK4OQQ2\_0&fm=neo%2Fmerchandising&iid=311b8b26-ad85-486d-b99c-830d8ede9681.SHOF4HFMJY73FT5E.SEARCH&ppt=Store+Browse&ppn=Store&ssid=nrathtsjy80000001533667797259&affid=nathgopin')

html = response.read()

soup = BeautifulSoup(html,"html5lib")

text = soup.get\_text(strip=True)

hi = soup.select('.\_35KyD6')

print(hi[0].getText())

The Output:



1. PRODUCT PRICE

The Code:

from nltk.tokenize import word\_tokenize

from nltk.corpus import stopwords

from bs4 import BeautifulSoup

import urllib.request

stop\_words = stopwords.words('english')

response = urllib.request.urlopen('https://www.flipkart.com/earton-sneakers-men/p/itmf4hynmfrpsawk?pid=SHOF4HFMJY73FT5E&lid=LSTSHOF4HFMJY73FT5EVXLELQ&marketplace=FLIPKART&srno=b\_1\_1&otracker=hp\_omu\_Deals+of+the+Day\_4\_Under+%E2%82%B9699%2BExtra+10%25\_IB7FUTK4OQQ2\_0&fm=neo%2Fmerchandising&iid=311b8b26-ad85-486d-b99c-830d8ede9681.SHOF4HFMJY73FT5E.SEARCH&ppt=Store+Browse&ppn=Store&ssid=nrathtsjy80000001533667797259&affid=nathgopin')

html = response.read()

soup = BeautifulSoup(html,"html5lib")

text = soup.get\_text(strip=True)

hi = soup.select('.\_1vC4OE.\_3qQ9m1')

print(hi[0].getText())

The Output:



1. PRODUCT DISCOUNT

The Code:

from nltk.tokenize import word\_tokenize

from nltk.corpus import stopwords

from bs4 import BeautifulSoup

import urllib.request

stop\_words = stopwords.words('english')

response = urllib.request.urlopen('https://www.flipkart.com/earton-sneakers-men/p/itmf4hynmfrpsawk?pid=SHOF4HFMJY73FT5E&lid=LSTSHOF4HFMJY73FT5EVXLELQ&marketplace=FLIPKART&srno=b\_1\_1&otracker=hp\_omu\_Deals+of+the+Day\_4\_Under+%E2%82%B9699%2BExtra+10%25\_IB7FUTK4OQQ2\_0&fm=neo%2Fmerchandising&iid=311b8b26-ad85-486d-b99c-830d8ede9681.SHOF4HFMJY73FT5E.SEARCH&ppt=Store+Browse&ppn=Store&ssid=nrathtsjy80000001533667797259&affid=nathgopin')

html = response.read()

soup = BeautifulSoup(html,"html5lib")

text = soup.get\_text(strip=True)

hi = soup.select('.VGWI6T.\_1iCvwn')

print(hi[0].getText()) <span>30% off</span>

<div class="VGWI6T \_1iCvwn"><span>30% off</span></div>

The Output:



1. PRODUCT IMAGE

The Code:

from nltk.tokenize import word\_tokenize

from nltk.corpus import stopwords

from bs4 import BeautifulSoup

import urllib.request

stop\_words = stopwords.words('english')

response = urllib.request.urlopen('https://www.flipkart.com/earton-sneakers-men/p/itmf4hynmfrpsawk?pid=SHOF4HFMJY73FT5E&lid=LSTSHOF4HFMJY73FT5EVXLELQ&marketplace=FLIPKART&srno=b\_1\_1&otracker=hp\_omu\_Deals+of+the+Day\_4\_Under+%E2%82%B9699%2BExtra+10%25\_IB7FUTK4OQQ2\_0&fm=neo%2Fmerchandising&iid=311b8b26-ad85-486d-b99c-830d8ede9681.SHOF4HFMJY73FT5E.SEARCH&ppt=Store+Browse&ppn=Store&ssid=nrathtsjy80000001533667797259&affid=nathgopin')

html = response.read()

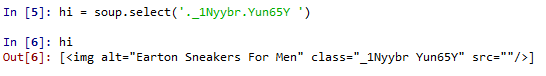
soup = BeautifulSoup(html,"html5lib")

text = soup.get\_text(strip=True)

hi = soup.select('.\_1Nyybr.Yun65Y ')

print(hi)

The Output:



Note: The Image is not saved to repository as the web is text scraped and not downloaded. Therefore image not shown. The actual image to be shown is :-

