```
!apt-get update
!apt install chromium-chromedriver
!pip install selenium

```

```
Hit:1 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
Hit:2 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease
Hit:3 http://security.ubuntu.com/ubuntu jammy-security InRelease
Ign:4 https://r2u.stat.illinois.edu/ubuntu jammy InRelease
Hit:5 http://archive.ubuntu.com/ubuntu jammy InRelease
Hit:6 https://r2u.stat.illinois.edu/ubuntu jammy Release
Hit:7 http://archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:8 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:9 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:10 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
Hit:12 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Reading package lists... Done
W: Skipping acquire of configured file 'main/source/Sources' as repository 'https://r2u.stat.illinois.edu/ubuntu jammy InRelease' d
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
chromium-chromedriver is already the newest version (1:85.0.4183.83-0ubuntu2.22.04.1).
0 upgraded, 0 newly installed, 0 to remove and 46 not upgraded.
Requirement already satisfied: selenium in /usr/local/lib/python3.10/dist-packages (4.23.0)
Requirement already satisfied: urllib3[socks]<3,>=1.26 in /usr/local/lib/python3.10/dist-packages (from selenium) (2.0.7)
Requirement already satisfied: trio~=0.17 in /usr/local/lib/python3.10/dist-packages (from selenium) (0.26.0)
Requirement already satisfied: trio-websocket~=0.9 in /usr/local/lib/python3.10/dist-packages (from selenium) (0.11.1)
Requirement already satisfied: certifi>=2021.10.8 in /usr/local/lib/python3.10/dist-packages (from selenium) (2024.7.4)
Requirement already satisfied: typing_extensions~=4.9.0 in /usr/local/lib/python3.10/dist-packages (from selenium) (4.9.0)
Requirement already satisfied: websocket-client==1.8.0 in /usr/local/lib/python3.10/dist-packages (from selenium) (1.8.0)
Requirement already satisfied: attrs>=23.2.0 in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (23.2.0)
Requirement already satisfied: sortedcontainers in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (2.4.0)
Requirement already satisfied: idna in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (3.7)
Requirement already satisfied: outcome in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (1.3.0.post0)
Requirement already satisfied: sniffio>=1.3.0 in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (1.3.1)
Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (1.2.2)
Requirement already satisfied: wsproto>=0.14 in /usr/local/lib/python3.10/dist-packages (from trio-websocket~=0.9->selenium) (1.2.0
Requirement already satisfied: pysocks!=1.5.7,<2.0,>=1.5.6 in /usr/local/lib/python3.10/dist-packages (from urllib3[socks]<3,>=1.26
Requirement already satisfied: h11<1,>=0.9.0 in /usr/local/lib/python3.10/dist-packages (from wsproto>=0.14->trio-websocket~=0.9->se
```

```
import requests
from bs4 import BeautifulSoup as bs
import lxml
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
import time
import pandas as pd
import random

```

```
# prompt: webdriver chrome functions does not accept options parameter

import requests
from bs4 import BeautifulSoup as bs
import lxml
from selenium import webdriver
from selenium.webdriver.chrome.options import Options
import time
import pandas as pd
import random

!apt-get update
!apt install chromium-chromedriver
!pip install selenium

chrome_options = webdriver.ChromeOptions()
chrome_options.add_argument('--headless')
chrome_options.add_argument('--no-sandbox')
chrome_options.add_argument('--disable-dev-shm-usage')
driver = webdriver.Chrome(options=chrome_options)

```

```
Hit:1 http://security.ubuntu.com/ubuntu jammy-security InRelease
Hit:2 https://cloud.r-project.org/bin/linux/ubuntu jammy-cran40/ InRelease
Hit:3 https://developer.download.nvidia.com/compute/cuda/repos/ubuntu2204/x86_64  InRelease
Hit:4 http://archive.ubuntu.com/ubuntu jammy InRelease
Ign:5 https://r2u.stat.illinois.edu/ubuntu jammy InRelease
Hit:6 http://archive.ubuntu.com/ubuntu jammy-updates InRelease
Hit:7 https://r2u.stat.illinois.edu/ubuntu jammy Release
Hit:8 http://archive.ubuntu.com/ubuntu jammy-backports InRelease
Hit:9 https://ppa.launchpadcontent.net/deadsnakes/ppa/ubuntu jammy InRelease
Hit:10 https://ppa.launchpadcontent.net/graphics-drivers/ppa/ubuntu jammy InRelease
```

```
Hit:11 https://ppa.launchpadcontent.net/ubuntugis/ppa/ubuntu jammy InRelease
Reading package lists... Done
W: Skipping acquire of configured file 'main/source/Sources' as repository 'https://r2u.stat.illinois.edu/ubuntu jammy InRelease' d
Reading package lists... Done
Building dependency tree... Done
Reading state information... Done
chromium-chromedriver is already the newest version (1:85.0.4183.83-0ubuntu2.22.04.1).
0 upgraded, 0 newly installed, 0 to remove and 46 not upgraded.
Requirement already satisfied: selenium in /usr/local/lib/python3.10/dist-packages (4.23.0)
Requirement already satisfied: urllib3[socks]<3,>=1.26 in /usr/local/lib/python3.10/dist-packages (from selenium) (2.0.7)
Requirement already satisfied: trio~=0.17 in /usr/local/lib/python3.10/dist-packages (from selenium) (0.26.0)
Requirement already satisfied: trio-websocket~=0.9 in /usr/local/lib/python3.10/dist-packages (from selenium) (0.11.1)
Requirement already satisfied: certifi>=2021.10.8 in /usr/local/lib/python3.10/dist-packages (from selenium) (2024.7.4)
Requirement already satisfied: typing_extensions~=4.9.0 in /usr/local/lib/python3.10/dist-packages (from selenium) (4.9.0)
Requirement already satisfied: websocket-client==1.8.0 in /usr/local/lib/python3.10/dist-packages (from selenium) (1.8.0)
Requirement already satisfied: attrs>=23.2.0 in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (23.2.0)
Requirement already satisfied: sortedcontainers in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (2.4.0)
Requirement already satisfied: idna in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (3.7)
Requirement already satisfied: outcome in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (1.3.0.post0)
Requirement already satisfied: sniffio>=1.3.0 in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (1.3.1)
Requirement already satisfied: exceptiongroup in /usr/local/lib/python3.10/dist-packages (from trio~=0.17->selenium) (1.2.2)
Requirement already satisfied: wsproto>=0.14 in /usr/local/lib/python3.10/dist-packages (from trio-websocket~=0.9->selenium) (1.2.0
Requirement already satisfied: pysocks!=1.5.7,<2.0,>=1.5.6 in /usr/local/lib/python3.10/dist-packages (from urllib3[socks]<3,>=1.26
Requirement already satisfied: h11<1,>=0.9.0 in /usr/local/lib/python3.10/dist-packages (from wsproto>=0.14->trio-websocket~=0.9->se
```

Getting links of each product

```python
keyword = 'laptop'
page = int(input('enter the number of pages you want to scrape '))
links = []
for page_no in range(1,int(page+1)):
    options = Options()
    options.add_argument('--headless')
    options.add_argument('--disable-gpu')
    options.add_argument('--no-sandbox')
    options.add_argument('--disable-dev-shm-usage')
    driver = webdriver.Chrome(options=options)
    #headers = {'user-agent':user_agent}
    url =f"https://www.amazon.in/s?k={keyword}&page={page_no}&qid=1638731354&ref=sr_pg_{page_no}"
    driver.get(url)
    time.sleep(1)

    source_code = driver.page_source
    driver.quit()
    response = requests.get(url).content
    soup = bs(source_code,'lxml')
    all_product_soup = soup.findAll('a',attrs={"class":"a-link-normal s-underline-text s-underline-link-text s-link-style a-text-nor
    for x in all_product_soup:
        links.append(x.attrs['href'] )
```

```
enter the number of pages you want to scrape 25
```

```python
links
```

```
 '/ASUS-Vivobook-IntelCore-Fingerprint-X1404ZA-NK322WS/dp/B0CCPBPYSB/ref=sr_1_36?
dib=eyJ2IjoiMSJ9.ywKU3HETe2S_zm3GYX_oHCBDBwPF60k77nr-kMEDOpDhgbxTgxQGXve2NVQimWOLiZZPrtC4r24S0ZeS8-
HS_y59kjgfUX59UUvmW_Fyg4adPQzv1nO8eF4Xg_3if_35seWaWl1cP0GFEaX3V1I29yGCJsaDJGAOi2d82rSCKha40zNLxmamekfq-
gK12ilbnCSMYn_qQG1sj3_R5PGkehSOIarPy5nH3gG9VdDRnz8.tX-
UNEuyOWZbFvkouhjLpMd05TZlL0eqruPVxH4U1vc&dib_tag=se&keywords=laptop&qid=1721749184&sr=8-36',
 '/Lenovo-Smartchoice-i5-12450HX-Graphics-83GS003UIN/dp/B0CX8XPKWJ/ref=sr_1_37?
dib=eyJ2IjoiMSJ9.ywKU3HETe2S_zm3GYX_oHCBDBwPF60k77nr-kMEDOpDhgbxTgxQGXve2NVQimWOLiZZPrtC4r24S0ZeS8-
HS_y59kjgfUX59UUvmW_Fyg4adPQzv1nO8eF4Xg_3if_35seWaWl1cP0GFEaX3V1I29yGCJsaDJGAOi2d82rSCKha40zNLxmamekfq-
gK12ilbnCSMYn_qQG1sj3_R5PGkehSOIarPy5nH3gG9VdDRnz8.tX-
UNEuyOWZbFvkouhjLpMd05TZlL0eqruPVxH4U1vc&dib_tag=se&keywords=laptop&qid=1721749184&sr=8-37',
 '/HP-i3-1315U-14-inch-Graphics-gr0000TU/dp/B0C9DL7THT/ref=sr_1_38?dib=eyJ2IjoiMSJ9.ywKU3HETe2S_zm3GYX_oHCBDBwPF60k77nr-
kMEDOpDhgbxTgxQGXve2NVQimWOLiZZPrtC4r24S0ZeS8-
HS_y59kjgfUX59UUvmW_Fyg4adPQzv1nO8eF4Xg_3if_35seWaWl1cP0GFEaX3V1I29yGCJsaDJGAOi2d82rSCKha40zNLxmamekfq-
gK12ilbnCSMYn_qQG1sj3_R5PGkehSOIarPy5nH3gG9VdDRnz8.tX-
UNEuyOWZbFvkouhjLpMd05TZlL0eqruPVxH4U1vc&dib_tag=se&keywords=laptop&qid=1721749184&sr=8-38',
 '/HP-i3-1215U-Anti-Glare-Speakers-15s-fq5326TU/dp/B0CYQ257C1/ref=sr_1_39?dib=eyJ2IjoiMSJ9.ywKU3HETe2S_zm3GYX_oHCBDBwPF60k77nr-
kMEDOpDhgbxTgxQGXve2NVQimWOLiZZPrtC4r24S0ZeS8-
HS_y59kjgfUX59UUvmW_Fyg4adPQzv1nO8eF4Xg_3if_35seWaWl1cP0GFEaX3V1I29yGCJsaDJGAOi2d82rSCKha40zNLxmamekfq-
gK12ilbnCSMYn_qQG1sj3_R5PGkehSOIarPy5nH3gG9VdDRnz8.tX-
UNEuyOWZbFvkouhjLpMd05TZlL0eqruPVxH4U1vc&dib_tag=se&keywords=laptop&qid=1721749184&sr=8-39',
 '/Refurbished-Lenovo-ThinkPad-Windows-Graphics/dp/B0CR49QPM9/ref=sr_1_40?dib=eyJ2IjoiMSJ9.ywKU3HETe2S_zm3GYX_oHCBDBwPF60k77nr-
kMEDOpDhgbxTgxQGXve2NVQimWOLiZZPrtC4r24S0ZeS8-
HS_y59kjgfUX59UUvmW_Fyg4adPQzv1nO8eF4Xg_3if_35seWaWl1cP0GFEaX3V1I29yGCJsaDJGAOi2d82rSCKha40zNLxmamekfq-
gK12ilbnCSMYn_qQG1sj3_R5PGkehSOIarPy5nH3gG9VdDRnz8.tX-
UNEuyOWZbFvkouhjLpMd05TZlL0eqruPVxH4U1vc&dib_tag=se&keywords=laptop&qid=1721749184&sr=8-40',
 '/Dell-R5-5500U-35-56cm-Spill-Resistant-Keyboard/dp/B0CLH81BGH/ref=sr_1_41?
dib=eyJ2IjoiMSJ9.ywKU3HETe2S_zm3GYX_oHCBDBwPF60k77nr-kMEDOpDhgbxTgxQGXve2NVQimWOLiZZPrtC4r24S0ZeS8-
HS_y59kjgfUX59UUvmW_Fyg4adPQzv1nO8eF4Xg_3if_35seWaWl1cP0GFEaX3V1I29yGCJsaDJGAOi2d82rSCKha40zNLxmamekfq-
```

```

import re

#text = '/HP-i3-1215U-15-6-inch-Anti-Glare-15s-fy5006TU/dp/B0CJBL2QWY/ref=sr_1_48?dib=eyJ2IjoiMSJ9.ywKU3HETe2S_zm3GYX_oHCBDBwPF60k77

updated_links=[]
for text in links:
  pattern = r'^(.*?/dp/[A-Z0-9]+)'

  match = re.match(pattern, text)
  if match:
    extracted_text = match.group(1)
    updated_links.append("https://www.amazon.in"+extracted_text)


updated_links
```

```
['https://www.amazon.in/Lenovo-Microsoft-Lifetime-Dual-core-Ethernet/dp/B0CR8VHCSV',
 'https://www.amazon.in/Lenovo-IdeaPad-39-62cm-300nits-82K20289IN/dp/B0CL5L59Z9',
 'https://www.amazon.in/Refurbished-Lenovo-ThinkPad-Windows-Graphics/dp/B0CR496TBN',
 'https://www.amazon.in/HP-i5-1334U-15-6-inch-graphics-speakers/dp/B0CTKHG3F6',
 'https://www.amazon.in/Acer-i5-1235U-Windows-Graphics-AL15-52/dp/B0CLTW4D7T',
 'https://www.amazon.in/Lenovo-IdeaPad-39-62cm-Warranty-82RK006DIN/dp/B0B4JPC8GT',
 'https://www.amazon.in/Apple-MacBook-Chip-13-inch-256GB/dp/B08N5W4NNB',
 'https://www.amazon.in/HP-i5-1235U-15-6-inch-Anti-Glare-15s-fy5007TU/dp/B0CJBP38HR',
 'https://www.amazon.in/Refurbished-HP-Chromebook-Bluetooth-Graphics/dp/B0D335SPLQ',
 'https://www.amazon.in/HP-i3-1215U-Graphics-Speakers-dy5008TU/dp/B0BZS88YPT',
 'https://www.amazon.in/Lenovo-IdeaPad-39-62cm-Windows-82R400BGIN/dp/B09MM58Y7Q',
 'https://www.amazon.in/HP-15-6-inch-Graphics-Speakers-ey2001AU/dp/B0C9DDN2JB',
 'https://www.amazon.in/HP-i5-1235U-15-6-inch-graphics-fy5008TU/dp/B0D4M1XZ3D',
 'https://www.amazon.in/HP-i3-1315U-15-6-inch-Graphics-Speakers/dp/B0C3RF3HT3',
 'https://www.amazon.in/Dell-Laptop-i5-1235U-Processor-Spill-Resistant/dp/B0D2DNCMB4',
 'https://www.amazon.in/Dell-Smartchoice-G15-5530-Gaming-i5-13450HX/dp/B0CRKXDX83',
 'https://www.amazon.in/Dell-Inspiron-3530-i3-1305U-Comfortview/dp/B0C4ZM63RP',
 'https://www.amazon.in/Refurbished-Dell-Latitude-Windows-Graphics/dp/B0D8BGMLGH',
 'https://www.amazon.in/Lenovo-IdeaPad-i7-13620H-38-1cm-83EM008GIN/dp/B0D6NCVQZQ',
 'https://www.amazon.in/ASUS-Vivobook-IntelCore-Fingerprint-X1404ZA-NK322WS/dp/B0CCPBPYSB',
 'https://www.amazon.in/Lenovo-Smartchoice-i5-12450HX-Graphics-83GS003UIN/dp/B0CX8XPKWJ',
 'https://www.amazon.in/HP-i3-1315U-14-inch-Graphics-gr0000TU/dp/B0C9DL7THT',
 'https://www.amazon.in/HP-i3-1215U-Anti-Glare-Speakers-15s-fq5326TU/dp/B0CYQ257C1',
 'https://www.amazon.in/Refurbished-Lenovo-ThinkPad-Windows-Graphics/dp/B0CR49QPM9',
 'https://www.amazon.in/Dell-R5-5500U-35-56cm-Spill-Resistant-Keyboard/dp/B0CLH81BGH',
 'https://www.amazon.in/Acer-Predator-Processor-Windows-PHN16-72/dp/B0CXPXW4VY',
 'https://www.amazon.in/HP-i3-1215U-15-6-inch-graphics-speakers/dp/B0D4LZMJ5Z',
 'https://www.amazon.in/ASUS-15-6-inch-Integrated-Transparent-X515MA-BR011W/dp/B09SGGB687',
 'https://www.amazon.in/Acer-Smartchoice-Premium-Windows-AL15-41/dp/B0CWTSF1TK',
 'https://www.amazon.in/Refurbished-Lenovo-ThinkPad-Windows-Graphics/dp/B0CR495DW5',
 'https://www.amazon.in/ASUS-Vivobook-IntelCore-Fingerprint-X1404ZA-NK321WS/dp/B0CCP9PH92',
 'https://www.amazon.in/HP-i3-1215U-15-6-inch-Anti-Glare-15s-fy5006TU/dp/B0CJBL2QWY']
```

extracting product title details by inspecting html cde

```python

all_details = []
for link in updated_links:
  options = Options()
  options.add_argument('--headless')
  options.add_argument('--disable-gpu')
  options.add_argument('--no-sandbox')
  options.add_argument('--disable-dev-shm-usage')
  driver = webdriver.Chrome(options=options)
  driver.get(link)
  time.sleep(2)
  source = driver.page_source
  soup = bs(source,'lxml')

  # Find title
  title_span = soup.find('span', {'id': 'productTitle'})
  if title_span:
    title = title_span.text.strip()
    all_details.append(title)
  else:
    all_details.append("Title not found")
```

```python
all_details
```

```
['Title not found',
 'Title not found',
 'Title not found',
 'HP Laptop 15, 13th Gen Intel Core i5-1334U, 15.6-inch (39.6 cm), FHD, 16GB DDR4, 512GB SSD, Intel Iris Xe graphics, Backlit KB,
MSO, Dual speakers (Win 11, Silver, 1.59 kg), fd0221TU',
 'Title not found',
 'Title not found',
 'Apple MacBook Air Laptop M1 chip, 13.3-inch/33.74 cm Retina Display, 8GB RAM, 256GB SSD Storage, Backlit Keyboard, FaceTime HD
Camera, Touch ID. Works with iPhone/iPad; Space Grey',
 'Title not found',
 'Title not found',
 'HP Laptop 14s, 12th Gen Intel Core i3-1215U, 14-inch (35.6 cm), FHD, 8GB DDR4, 512GB SSD, Intel UHD graphics, Thin & light, Dual
speakers (Win 11, MSO 2021, Silver, 1.46 kg), dy5008TU',
 'Title not found',
 'Title not found',
 'Title not found',
 'Title not found',
 'Title not found',
 'Title not found',
 'Dell Inspiron 3530 Thin & Light Laptop, 13th Gen Intel Core i3-1305U/8GB/512GB SSD/15.6" (39.62cm) 120Hz Refresh Rate on a FHD
IPS Display/Windows 11 + MSO\'21+McAfee 15 Month/Carbon Black/1.62kg',
 'Title not found',
 'Lenovo IdeaPad Slim 3 13th Gen Intel Core i7-13620H 15" (38.1cm) FHD IPS 300 Nits Thin & Light Laptop (16GB/512GB SSD/Win 11/MSO
21/1Yr ADP Free/Alexa built-in/3 mon Game Pass/Grey/1.6Kg), 83EM008GIN',
 'Title not found',
 'Title not found',
 'HP Laptop 14, 13th Gen Intel Core i3-1315U, 14-inch (35.6 cm), FHD, 8GB DDR4, 512GB SSD, Intel UHD Graphics, FHD Camera w/Privacy
Shutter, Thin & Light (Win 11, MSO 2021, Blue, 1.4 kg), gr0000TU',
 'Title not found',
 'Title not found',
 'Title not found',
 'Title not found',
 'Title not found',
 'ASUS VivoBook 15 (2021) Thin and Light Laptop, Dual Core Intel Celeron N4020, 15.6-inch (39.62 cm) HD, (4GB RAM/256GB
SSD/Integrated Graphics/Windows 11 Home/Transparent Silver/1.8 Kg), X515MA-BR011W',
 'Title not found',
 'Title not found',
 'ASUS Vivobook 14 Thin and Light Laptop, IntelCore i3-1215U 12th Gen, 14" (35.56 cm) FHD, (8 GB RAM/512GB SSD/Win11/Office
2021/Fingerprint/42WHr /Blue/1.40 kg), X1404ZA-NK321WS',
 'Title not found']
```

Getting Product Details

```
#use of headless selenium to get the dynamic content from amazon product page
data =[]
for product_link in updated_links:
    options = Options()
    options.add_argument('--headless')
    options.add_argument('--disable-gpu')
    options.add_argument('--no-sandbox')
    options.add_argument('--disable-dev-shm-usage')
    driver = webdriver.Chrome(options=options)
    driver.get(product_link)
    time.sleep(1)

    source_code = driver.page_source
    driver.quit()
    product_soup =  bs(source_code,'lxml')
    product_details={}
    for details in product_soup.find_all('table',attrs={'id':'productDetails_techSpec_section_1'}):
        product_details[details.th.text.rstrip()]=details.td.text.lstrip()

    data.append(product_details)
```

```
df1=pd.DataFrame(all_details)
```

```
df1.tail()
```

	0
27	ASUS VivoBook 15 (2021) Thin and Light Laptop,...
28	Title not found
29	Title not found
30	ASUS Vivobook 14 Thin and Light Laptop, IntelC...
31	Title not found

```
# keep on extracting with different ids
```