Build a web content extractor that extracts text from a HTML file.

Solution:

In [27]:
```
import requests
import bs4
import html5lib
```

Get all the URLs from site:

In [28]:
```
def URL_from_site(url, *args):
    res = requests.get(url)
    res.raise_for_status()
    amritaSoup = bs4.BeautifulSoup(res.text,"lxml")
    type(amritaSoup)

    file = open("Amazonfile.txt","w") # This file save all URLs to Amazonfile.txt file

    for a in amritaSoup.find_all('a',href=True):
        file.writelines('URL: '+ str(url)+ a['href']+'\n')

    file.close()

    file_in = open("Amazonfile.txt","r")  # Line reads all URLs from Amazonfile.txt file

    if len(args) == 0:
        for line in file_in:

            print(line)
        return

    page = []

    if len(args) != 0:
        for line in file_in:
            if (args[0]) in line:
#                 print(line)
                page.append(line)
        return page
```

Print HTML content:

In [29]:
```
def HTML_content(url):

    url= mobile_url[0].strip('URL: \n')
    r = requests.get(url)

    soup = bs4.BeautifulSoup(r.content, 'html5lib')
    print(soup.prettify())

    return
```

Get HTML text:

In [30]:
```
def HtML_2_Text_converter(url):

    url= mobile_url[0].strip('URL: \n')

    req = requests.get(url)
    html_page = req.text

    soup = bs4.BeautifulSoup(html_page, "html.parser")

    html_text = soup.get_text()

    Text = []

    for text in html_text:
        for word in text:
            if '\n' in word:
                break
        else: Text.append(word)

    return html_text, "".join(Text)
```

Get URLs from https://www.amazon.in/ (https://www.amazon.in/)

In [36]:
```
url = 'https://www.amazon.in/'
URL_from_site(url)
```

```
URL: https://www.amazon.in//ref=nav_logo (https://www.amazon.in//ref=nav_logo)

URL: https://www.amazon.in//gp/customer-preferences/select-language/ref=topnav_lang?preferencesReturnUrl=%2F (https://www.amazon.in//gp/customer-preferences/select-language/ref=topnav_lang?preferencesReturnUrl=%2F)

URL: https://www.amazon.in/https://www.amazon.in/ap/signin?openid.pape.max_auth_age=0&openid.return_to=https%3A%2F%2Fwww.amazon.in%2F%3F_encoding%3DUTF8%26ref_%3Dnav_ya_signin&openid.identity=http%3A%2F%2Fspecs.openid.net%2Fauth%2F2.0%2Fidentifier_select&openid.assoc_handle=inflex&openid.mode=checkid_setup&openid.claimed_id=http%3A%2F%2Fspecs.openid.net%2Fauth%2F2.0%2Fidentifier_select&openid.ns=http%3A%2F%2Fspecs.openid.net%2Fauth%2F2.0& (https://www.amazon.in/https://www.amazon.in/ap/signin?openid.pape.max_auth_age=0&openid.return_to=https%3A%2F%2Fwww.amazon.in%2F%3F_encoding%3DUTF8%26ref_%3Dnav_ya_signin&openid.identity=http%3A%2F%2Fspecs.openid.net%2Fauth%2F2.0%2Fidentifier_select&openid.assoc_handle=inflex&openid.mode=checkid_setup&openid.claimed_id=http%3A%2F%2Fspecs.openid.net%2Fauth%2F2.0%2Fidentifier_select&openid.ns=http%3A%2F%2Fspecs.openid.net%2Fauth%2F2.0&)

URL: https://www.amazon.in//gp/css/order-history?ref_=nav_orders_first (https://www.amazon.in//gp/css/order-history?ref_=nav_orders_first)
```

Get URL related to an item (mobile):

In [37]:
```
mobile_url = URL_from_site(url, 'mobile')
mobile_url
```

Out[37]:
```
['URL: https://www.amazon.in//mobile-phones/b/?ie=UTF8&node=1389401031&ref_=nav_cs_mobiles_9292c6cb7b394d30b2467b8f631090a7\n'] (https://www.amazon.in//mobile-phones/b/?ie=UTF8&node=1389401031&ref_=nav_cs_mobiles_9292c6cb7b394d30b2467b8f631090a7\n'])
```

Get HTML content of 'mobile_url':

In [38]:
```
HTML_content(mobile_url)
```

```
<!DOCTYPE html>
<html class="a-no-js" data-19ax5a9jf="dingo" lang="en-in">
 <!-- sp:feature:head-start -->
 <head>
  <script>
   var aPageStart = (new Date()).getTime();
  </script>
  <meta charset="utf-8"/>
  <!-- sp:feature:cs-optimization -->
  <meta content="on" http-equiv="x-dns-prefetch-control"/>
  <link href="https://images-eu.ssl-images-amazon.com" rel="dns-prefetch"/>
  <link href="https://m.media-amazon.com" rel="dns-prefetch"/>
  <link href="https://completion.amazon.com" rel="dns-prefetch"/>
  <!-- sp:feature:aui-assets -->
  <link href="https://images-eu.ssl-images-amazon.com/images/I/11EIQ5IGqaL._RC|012LjolmrML.css,41
DO8IyHTdL.css,21qPwhPKAAL.css,01Vctty9pOL.css,017DsKjNQJL.css,0131vqwP5UL.css,41EWOOlBJ9L.css,11g
KzVUTNZL.css,01ElnPiDxWL.css,11bGSgD5pDL.css,01Dm5eKVxwL.css,01IdKcBuAdL.css,01y-XAlI+2L.css,01Zf
XnjPmmL.css,01oDR3IULNL.css,31q1y1irc5L.css,01XPHJk60-L.css,01R0k0yxPXL.css,21xVR0NtxzL.css,11gne
A3MtJL.css,21fecG8pUzL.css,01RddH8vm-L.css,01CFUgsA-YL.css,21AmhU6t0sL.css,11zGrJZ9D2L.css,11tRp6
```

Get HTML text of 'mobile_url':

In [39]:
```
html_text,  Text = HtML_2_Text_converter(mobile_url)
Text
```

Out[39]:
```
"Mobile Phones: Buy New Mobiles Online at Best Prices in India | Buy Cell Phones Online - Amazon.inS
kip to main content.in                    Hello                                        Select your address
Mobiles & AccessoriesSelect the department you want to search inMobiles & AccessoriesAll CategoriesD
ealsAlexa SkillsAmazon DevicesAmazon FashionAmazon FreshAmazon PantryAppliancesApps & GamesBabyBeaut
yBooksCar & MotorbikeClothing & AccessoriesCollectiblesComputers & AccessoriesElectronicsFurnitureGa
rden & OutdoorsGift CardsGrocery & Gourmet FoodsHealth & Personal CareHome & KitchenIndustrial & Sci
entificJewelleryKindle StoreLuggage & BagsLuxury BeautyMovies & TV ShowsMusicMusical InstrumentsOffi
ce ProductsPet SuppliesPrime VideoShoes & HandbagsSoftwareSports, Fitness & OutdoorsTools & Home Imp
rovementToys & GamesUnder ₹500Video GamesWatchesHello, Sign inAccount & Lists    Account    Returns&
Orders        Cart        AllBest SellersMobilesToday's DealsFashionNew ReleasesPrime Electronics Cu
stomer ServiceAmazon PayHome & KitchenComputersBooksToys & GamesSellBeauty & Personal CareCar & Moto
rbikeGift CardsGrocery & Gourmet FoodsHealth, Household & Personal CareSports, Fitness & OutdoorsGif
t Ideas\tBabyVideo GamesPet SuppliesAmazonBasicsCouponsKindle eBooksHome ImprovementSubscribe & Save
Mobiles & Accessories                Laptops & Accessories                TV & Home Entertainment
Audio                Cameras                Computer Peripherals                Smart Technology
Musical Instruments                Office & Stationery          https://www.amazon.in/gp/goldbox/?pf
_rd_p=52e6a193-83f2-4e5b-acbb-6d28c918c36b&pf_rd_s=merchandised-search-4&pf_rd_t=101&pf_rd_i=1389401
031&pf_rd_m=A1VBAL9TL5WCBF&pf_rd_r=XCCJMT5YH2TK8DP9BNTV&pf_rd_r=XCCJMT5YH2TK8DP9BNTV&pf_rd_p=52e6a19
3-83f2-4e5b-acbb-6d28c918c36bhttps://www.amazon.in/gp/goldbox/?pf_rd_p=7271e5cd-bd21-4eb6-8940-7ce00
6673c0d&pf_rd_s=merchandised-search-18&pf_rd_t=101&pf_rd_i=1389401031&pf_rd_m=A1VBAL9TL5WCBF&pf_rd_r
=XCCJMT5YH2TK8DP9BNTV&pf_rd_r=XCCJMT5YH2TK8DP9BNTV&pf_rd_p=7271e5cd-bd21-4eb6-8940-7ce006673c0d1-12
 (https://www.amazon.in/gp/goldbox/?pf_rd_p=52e6a193-83f2-4e5b-acbb-6d28c918c36b&pf_rd_s=merchandise
d-search-4&pf_rd_t=101&pf_rd_i=1389401031&pf_rd_m=A1VBAL9TL5WCBF&pf_rd_r=XCCJMT5YH2TK8DP9BNTV&pf_rd_
r=XCCJMT5YH2TK8DP9BNTV&pf_rd_p=52e6a193-83f2-4e5b-acbb-6d28c918c36bhttps://www.amazon.in/gp/goldbo
x/?pf_rd_p=7271e5cd-bd21-4eb6-8940-7ce006673c0d&pf_rd_s=merchandised-search-18&pf_rd_t=101&pf_rd_i=1
389401031&pf_rd_m=A1VBAL9TL5WCBF&pf_rd_r=XCCJMT5YH2TK8DP9BNTV&pf_rd_r=XCCJMT5YH2TK8DP9BNTV&pf_rd_p=7
271e5cd-bd21-4eb6-8940-7ce006673c0d1-12) of over 70,000 results forMobiles & AccessoriesBest selleri
n In-Ear Headphones    boAt Bassheads 100 in Ear Wired Earphones with Mic(Black)by Boat123,537₹379.0
0₹379.00₹999.00₹999.00FulfilledFREE Delivery on orders over ₹499.00.Details                    D
elivery by: Wednesday, March 24Details    Redmi 9A (Nature Green, 2GB Ram, 32GB Storage) | 2GHz Octa
-core Helio G25 Processorby Redmi20,691₹6,799.00₹6,799.00₹8,499.00₹8,499.00FulfilledFREE Delivery.De
tailsIn stock on March 27, 2021.Best sellerin Electronics    Redmi 9 (Sky Blue, 4GB RAM, 64GB Storag
e)by Redmi17,207₹8,799.00₹8,799.00₹10,999.00₹10,999.00FulfilledFREE Delivery.Details
Delivery by: Wednesday, March 24Details    Samsung Galaxy M12 (Blue,4GB RAM, 64GB Storage) 6000 mAh
 with 8nm Processor | True 48 MP Quad Camera | 90Hz Refresh Rateby Samsung24₹10,999.00₹10,999.00₹12,
999.00₹12,999.00FulfilledFREE Delivery.Details                    Delivery by: Wednesday, March
 24Details    Redmi 9A (Midnight Black, 2GB RAM, 32GB Storage) | 2GHz Octa-core Helio G25 Processorb
y Redmi20,691₹6,799.00₹6,799.00₹8,499.00₹8,499.00FulfilledFREE Delivery.DetailsIn stock on May 2, 20
21.    OnePlus Bullets Wireless Z in-Ear Bluetooth Earphones with Mic (Black)by OnePlus68,545₹1,999.
00₹1,999.00₹2,190.00₹2,190.00FulfilledFREE Delivery.Details                    Delivery by: Wedn
esday, March 24Details    OnePlus Bullets Wireless Z Bass Edition (Reverb Red)by OnePlus68,545₹1,99
9.00₹1,999.00₹2,190.00₹2,190.00FulfilledFREE Delivery.Details                    Delivery by: We
dnesday, March 24DetailsBest sellerin Smart Watches & Accessories    Noise Colorfit Pro 2 Full Touch
Control Smart Watch Jet Blackby Noise22,834₹2,999.00₹2,999.00₹4,999.00₹4,999.00FulfilledFREE Deliver
y.Details                    Delivery by: Wednesday, March 24Details    Boult Audio AirBass Z1 T
rue Wireless Earbuds with 24H Total Playtime, Touch Controls, IPX5 Sweatproof, Low Latency for Gamin
g and Voice Assistant(Black)by Boult Audio1₹1,599.00₹1,599.00₹5,999.00₹5,999.00FulfilledFREE Deliver
y.DetailsIn stock on March 26, 2021.    Redmi 9A (Sea Blue, 2GB Ram, 32GB Storage) | 2GHz Octa-core
 Helio G25 Processorby Redmi20,691₹6,799.00₹6,799.00₹8,499.00₹8,499.00FulfilledFREE Delivery.Details
In stock on April 30, 2021.    Redmi Note 10 (Shadow Black, 6GB RAM, 128GB Storage) - Super Amoled D
isplay | 48MP Sony Sensor IMX582 | Snapdragon 678 Processorby Redmi₹13,999.00₹13,999.00FulfilledFREE
Delivery.Details                    Delivery by: Wednesday, March 24Details    pTron Tangent Lit
e Bluetooth 5.0 Wireless Headphones with Hi-Fi Stereo Sound, 6Hrs Playtime, Lightweight Ergonomic Ne
ckband, Sweat-Resistant Magnetic Earbuds, Voice Assistant & Mic - (Black)by PTron7,364₹649.00₹649.00
₹1,800.00₹1,800.00FulfilledFREE Delivery.Details                    Delivery by: Wednesday, Marc
h 24DetailsSee all resultsAvg. Customer Review4 Stars & Up& Up3 Stars & Up& Up2 Stars & Up& Up1 Star
& Up& UpPay On DeliveryEligible for Pay On DeliveryAvailabilityInclude Out of StockNew ArrivalsLast
 30 daysLast 90 daysSellerAppario Retail Private LtdCloudtail IndiaKhurana CommunicationSpigen India
VagamaOye StuffShopMagicsTheGiftKartS.K RetailPuzzleStoreBrandsBoatRedmiSamsungOnePlusNoiseBoult Aud
ioPTronDiscount10% Off or more25% Off or more35% Off or more50% Off or moreMade for Amazon BrandsMad
e for AmazonDepartmentElectronicsMobiles & AccessoriesMobile AccessoriesSIM CardsSmartphones & Basic
MobilesItem ConditionNewRenewedUsedAmazon PrimeDealsToday's DealsPriceUnder ₹1,000₹1,000 - ₹5,000₹5,
000 - ₹10,000₹10,000 - ₹20,000Over ₹20,000 Unlimited FREE fast delivery, video streaming & more  Pri
me members enjoy unlimited free, fast delivery on eligible items, video streaming, ad-free music, ex
clusive access to deals & more.        >      Get startedBack to topGet to Know UsAbout UsCareersPres
s ReleasesAmazon CaresGift a SmileConnect with UsFacebookTwitterInstagramMake Money with UsSell on A
mazonSell under Amazon AcceleratorAmazon Global SellingBecome an AffiliateFulfilment by AmazonAdvert
ise Your ProductsAmazon Pay on MerchantsSee More Make Money with UsLet Us Help YouCOVID-19 and Amazo
nYour AccountReturns Centre100% Purchase ProtectionAmazon App DownloadAmazon Assistant DownloadHelpE
nglishChoose a language for shopping.AustraliaBrazilCanadaChinaFranceGermanyItalyJapanMexicoNetherla
ndsSingaporeSpainUnited Arab EmiratesUnited KingdomUnited StatesAbeBooks Books, art & collectiblesAm
azon Web Services Scalable Cloud Computing ServicesAudible Download Audio BooksDPReview Digital Phot
ographyIMDb Movies, TV & Celebrities\xa0Shopbop Designer Fashion Brands\tAmazon Business Everything
```

For Your BusinessPrime Now 2-Hour Delivery on Everyday ItemsAmazon Prime Music 70 million songs, ad-free\xa0Conditions of Use & SalePrivacy NoticeInterest-Based Ads© 1996-2021, Amazon.com, Inc. or its affiliates"