Chapter 2: Web Scraping, Applications and Tools

DR. LINDA MAHMOUDI

”يرفعِ اللهُ الذين أمنوا منكم والذين أوتوا العلم درجات“

Tools for Web Scraping

Web Scraping tools are specifically developed for extracting data from the internet. Also, known as web harvesting tools or data extraction tools, they are useful for anyone trying to collect specific data from websites as they provide the user with structured data extracting data from a number of websites. Some of the most popular Web Scraping tools are:

- Import.io
- Webhose.io
- Dexi.io
- Scrapinghub
- Parsehub

What tools can you use to scrape the web?

We’ve covered the basics of how to scrape the web for data, but how does this work from a technical standpoint? Often, web scraping requires some knowledge of programming languages, the most popular for the task being Python. Luckily, Python comes with a huge number of open-source libraries that make web scraping much easier. These include:

Python Libraries

BeautifulSoup: is Python library, commonly used to parse data from XML and HTML documents. Organizing this parsed content into more accessible trees, BeautifulSoup makes navigating and searching through large swathes of data much easier. It’s the go-to tool for many data analysts.

Scrapy: is a Python-based application framework that crawls and extracts structured data from the web. It’s commonly used for data mining, information processing, and for archiving historical content. As well as web scraping (which it was specifically designed for) it can be used as a general-purpose web crawler, or to extract data through APIs.

Python Libraries

Pandas: is another multi-purpose Python library used for data manipulation and indexing. It can be used to scrape the web in conjunction with BeautifulSoup. The main benefit of using pandas is that analysts can carry out the entire data analytics process using one language (avoiding the need to switch to other languages, such as R).

Parsehub

A bonus tool, in case you're not an experienced programmer!

Parsehub is a free online tool (to be clear, this one's not a Python library) that makes it easy to scrape online data. The only catch is that for full functionality you'll need to pay. But the free tool is worth playing around with, and the company offers excellent customer support.

How does a web scraper function?

So, we now know what web scraping is, and why different organizations use it. **But how does a web scraper work?** While the exact method differs depending on the software or tools you're using, all web scraping bots follow three basic principles:

Step 1: Making an HTTP request to a server

Step 2: Extracting and parsing (or breaking down) the website's code

Step 3: Saving the relevant data locally

Step 1: Making an HTTP request to a server

- As an individual, when you visit a website via your browser, you send what's called an HTTP request.
- This is basically the digital equivalent of knocking on the door, asking to come in. Once your request is approved, you can then access that site and all the information on it.
- Just like a person, a web scraper needs permission to access a site.
- Therefore, the first thing a web scraper does is send an HTTP request to the site they're targeting.

Step 2: Extracting and parsing the website's code

- Once a website gives a scraper access, the bot can read and extract the site's HTML or XML code.
- This code determines the website's content structure.
- The scraper will then parse the code (which basically means breaking it down into its constituent parts) so that it can identify and extract elements or objects that have been predefined by whoever set the bot loose! These might include specific text, ratings, classes, tags, IDs, or other information

Step 3: Saving the relevant data locally

- Once the HTML or XML has been accessed, scraped, and parsed, the web scraper will then store the relevant data locally.
- As mentioned, the data extracted is predefined by you (having told the bot what you want it to collect).
- Data is usually stored as structured data, often in an Excel file, such as a .csv or .xls format, mysql database.

- With these steps complete, you're ready to start using the data for your intended purposes. Easy, eh? And it's true…these three steps *do* make data scraping seem easy. In reality, though, the process isn't carried out just once, but countless times. This comes with its own swathe of problems that need solving.

- For instance, badly coded scrapers may send too many HTTP requests, which can crash a site. Every website also has different rules for what bots can and can't do. Executing web scraping code is just one part of a more involved process. Let's look at that now.

How to scrape the web (step-by-step)

Step one: Find the URLs you want to scrape

Step two: Inspect the page

Step three: Identify the data you want to extract

Step four: Write the necessary code

Step five: Execute the code

Step six: Storing the data

- Excel formats are the most common/ mysql database

amazon.in/gp/product/B09DMT6V7G/ref=s9_acss_bw_cg_SERIES_3a1_w?pf_rd_m=A1K21FY43GMZF8&pf_rd_s=merchandised-search-5&pf_rd_r=XMYNP8SWAGS...

amazon.in Hello Select your address | Computers & Accessories | Hello, Sign in Account & Lists | Returns & Orders | 0 Cart

All | Mobiles | Best Sellers | Electronics | Fashion | Customer Service | Prime | Amazon Pay | Today's Deals | Computers | Great Indian Festival | SHOP NOW

Electronics | Mobiles & Accessories | Laptops & Accessories | TV & Home Entertainment | Audio | Cameras | Computer Peripherals | Smart Technology | Musical Instruments | Office & Stationery

Computers & Accessories › Laptops › Traditional Laptops

Sponsored

Lenovo Ideapad Slim 3 10th Gen Intel Core i3 15.6" (39.62cm) FHD Thin & Light Laptop (8GB/256 GB SSD/UHD Graphics/Windows 10/MS Office/2 Year Warranty/Platinum Grey/1.7Kg), 81WB012DIN

Visit the Lenovo Store

4.0 out of 5 stars 456 ratings | 127 answered questions

M.R.P.: ~~₹52,290.00~~

Deal of the Day: ₹35,990.00

Ends in 11h 19m 41s

You Save: ₹16,300.00 (31%)

Inclusive of all taxes

FREE delivery: **Wednesday, Oct 13** Details

EMI starts at ₹1,694. No Cost EMI available EMI options

Save Extra with 4 offers

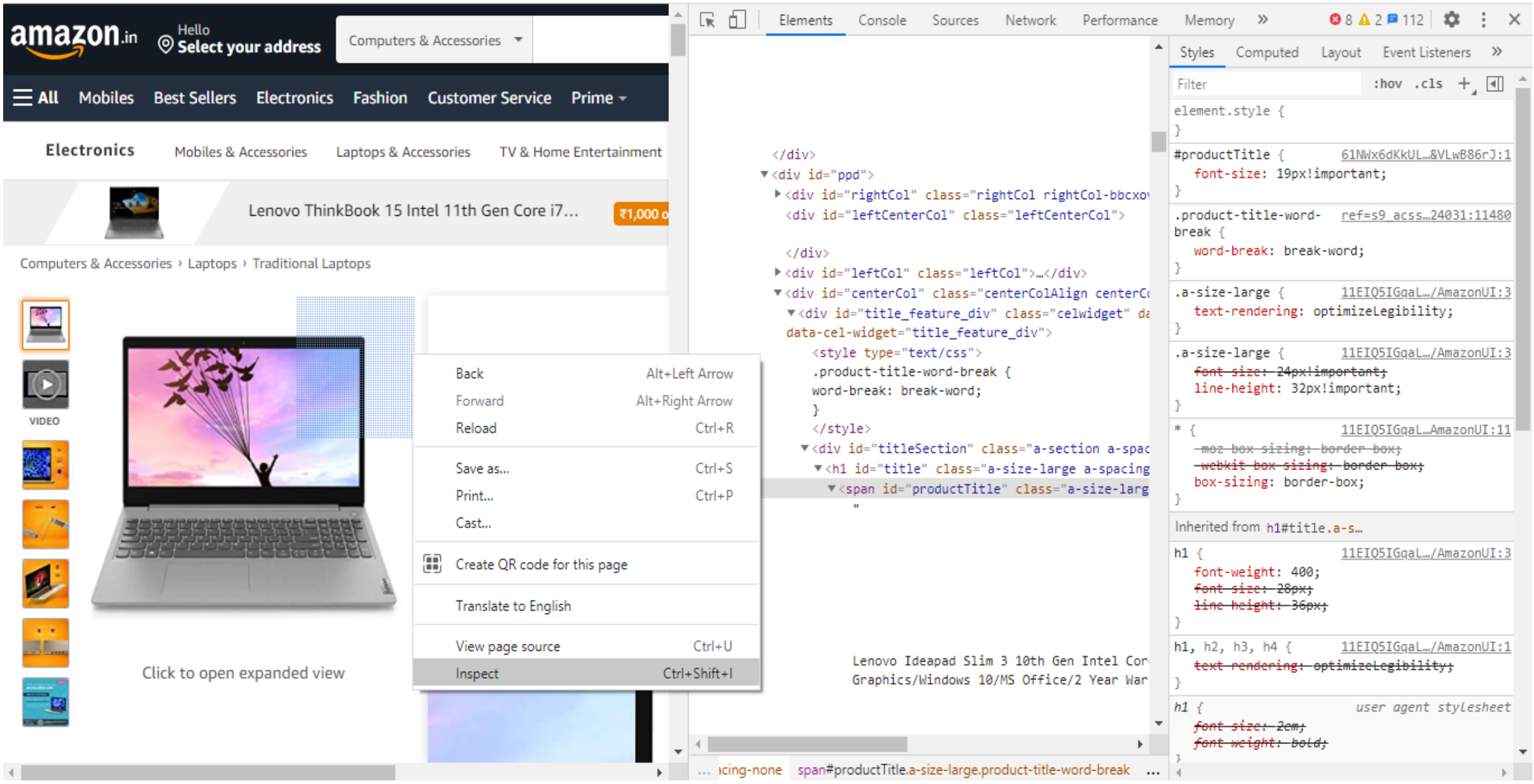
Share

With Exchange
Up to ₹ 18,050.00 off

Without Exchange
₹ 35,990.00 ~~₹ 52,290.00~~

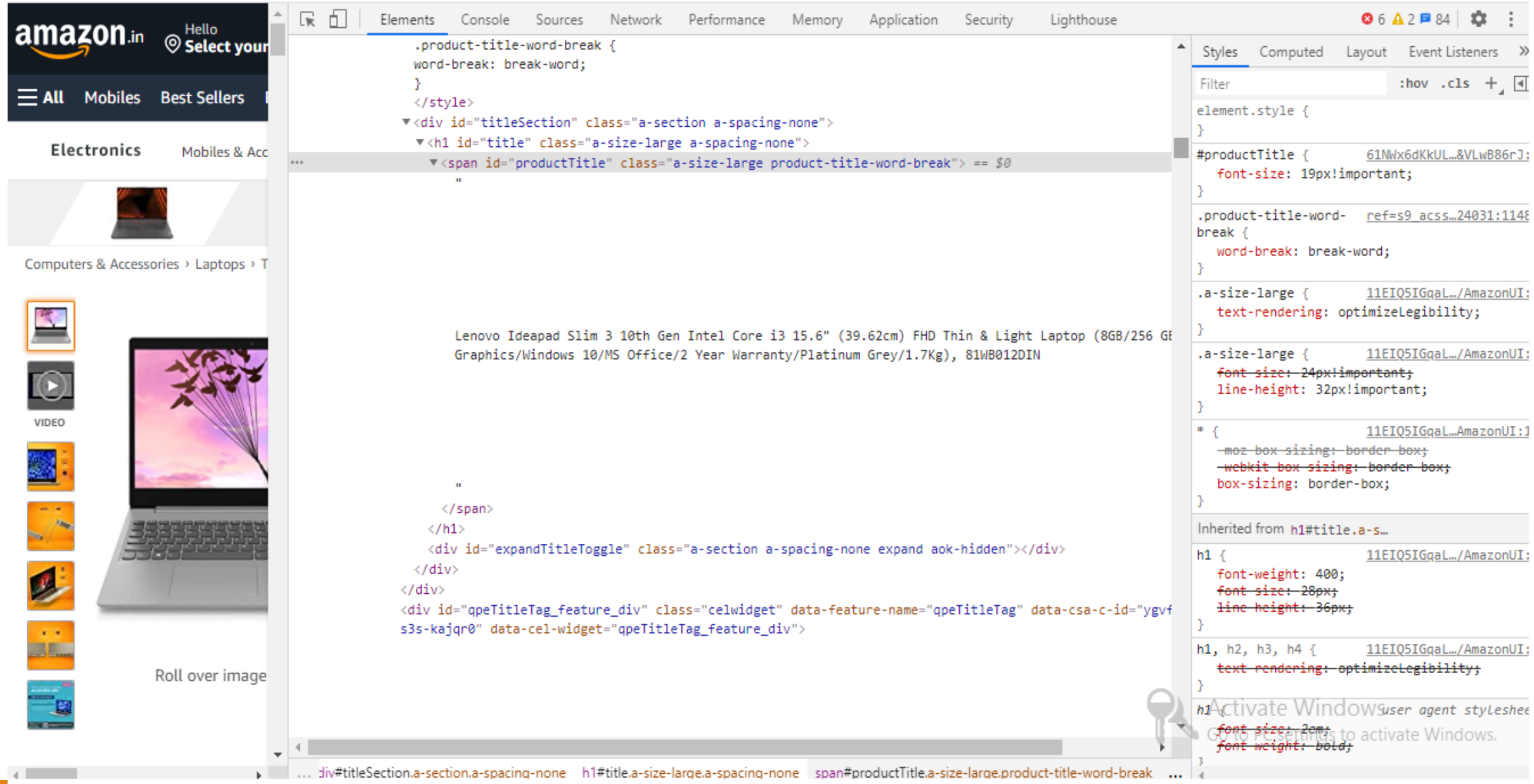
Add a Protection Plan:

- 1 Year Extended Warranty for ₹2,099.00
- 2 Year Care Plan (2 Year Extended Warranty Plan Post Expiry of 1st Year of Manufacturer Warranty for ₹1,241.00
- Quess Care Total Protection Plan for 1 Year Accidental Damage and Liquid Damage for Laptop Between 35001


amazon.in/gp/product/B09DMT6V7G/ref=s9_acss_bw_cg_SERIES_3a1_w?pf_rd_m=A1K21FY43GMZF8&pf_rd_s=merchandised-search-5&pf_rd_r=XMYNP8SWAGS...
amazon.in
Hello
Select your address
Computers & Accessories
All
Mobiles
Best Sellers
Electronics
Fashion
Customer Service
Prime
Electronics
Mobiles & Accessories
Laptops & Accessories
TV & Home Entertainment
Lenovo ThinkBook 15 Intel 11th Gen Core i7...
Computers & Accessories › Laptops › Traditional Laptops
VIDEO
Click to open expanded view
Back Alt+Left Arrow
Forward Alt+Right Arrow
Reload Ctrl+R
Save as... Ctrl+S
Print... Ctrl+P
Cast...
Create QR code for this page
Translate to English
View page source Ctrl+U
Inspect Ctrl+Shift+I
Elements
Console
Sources
Network
Performance
Memory
8 2 112
Styles
Computed
Layout
Event Listeners
Filter
:hov .cls
element.style {
}
#productTitle { 61NWx6dKkUL…&VLwB86rJ:1
font-size: 19px!important;
}
.product-title-word-break { ref=s9_acss…24031:11480
word-break: break-word;
}
.a-size-large { 11EIQ5IGqaL…/AmazonUI:3
text-rendering: optimizeLegibility;
}
.a-size-large { 11EIQ5IGqaL…/AmazonUI:3
font-size: 24px!important;
line-height: 32px!important;
}
* { 11EIQ5IGqaL…AmazonUI:11
-moz-box-sizing: border-box;
-webkit-box-sizing: border-box;
box-sizing: border-box;
}
Inherited from h1#title.a-s…
h1 { 11EIQ5IGqaL…/AmazonUI:3
font-weight: 400;
font-size: 28px;
line-height: 36px;
}
h1, h2, h3, h4 { 11EIQ5IGqaL…/AmazonUI:1
text-rendering: optimizeLegibility;
}
h1 { user agent stylesheet
font-size: 2em;
font-weight: bold;
}

<div id="rightCol" class="rightCol rightCol-bbcxo

…
<div id="centerCol" class="centerColAlign centerC
<div id="title_feature_div" class="celwidget" da
data-cel-widget="title_feature_div">
<style type="text/css">
.product-title-word-break {
word-break: break-word;
}
</style>
<div id="titleSection" class="a-section a-spac
<h1 id="title" class="a-size-large a-spacing
<span id="productTitle" class="a-size-larg
"
Lenovo Ideapad Slim 3 10th Gen Intel Cor
Graphics/Windows 10/MS Office/2 Year War
span#productTitle.a-size-large.product-title-word-break


amazon.in/gp/product/B09DMT6V7G/ref=s9_acss_bw_cg_SERIES_3a1_w?pf_rd_m=A1K21FY43GMZF8&pf_rd_s=merchandised-search-5&pf_rd_r=XMYNP8SWAGS...
amazon.in
Hello
Select your
All
Mobiles
Best Sellers
Electronics
Mobiles & Acc
Computers & Accessories › Laptops › T
VIDEO
Roll over image
Elements
Console
Sources
Network
Performance
Memory
Application
Security
Lighthouse
6 2 84
.product-title-word-break {
word-break: break-word;
}
</style>

<h1 id="title" class="a-size-large a-spacing-none">
 == $0
"
Lenovo Ideapad Slim 3 10th Gen Intel Core i3 15.6" (39.62cm) FHD Thin & Light Laptop (8GB/256 GB
Graphics/Windows 10/MS Office/2 Year Warranty/Platinum Grey/1.7Kg), 81WB012DIN
"

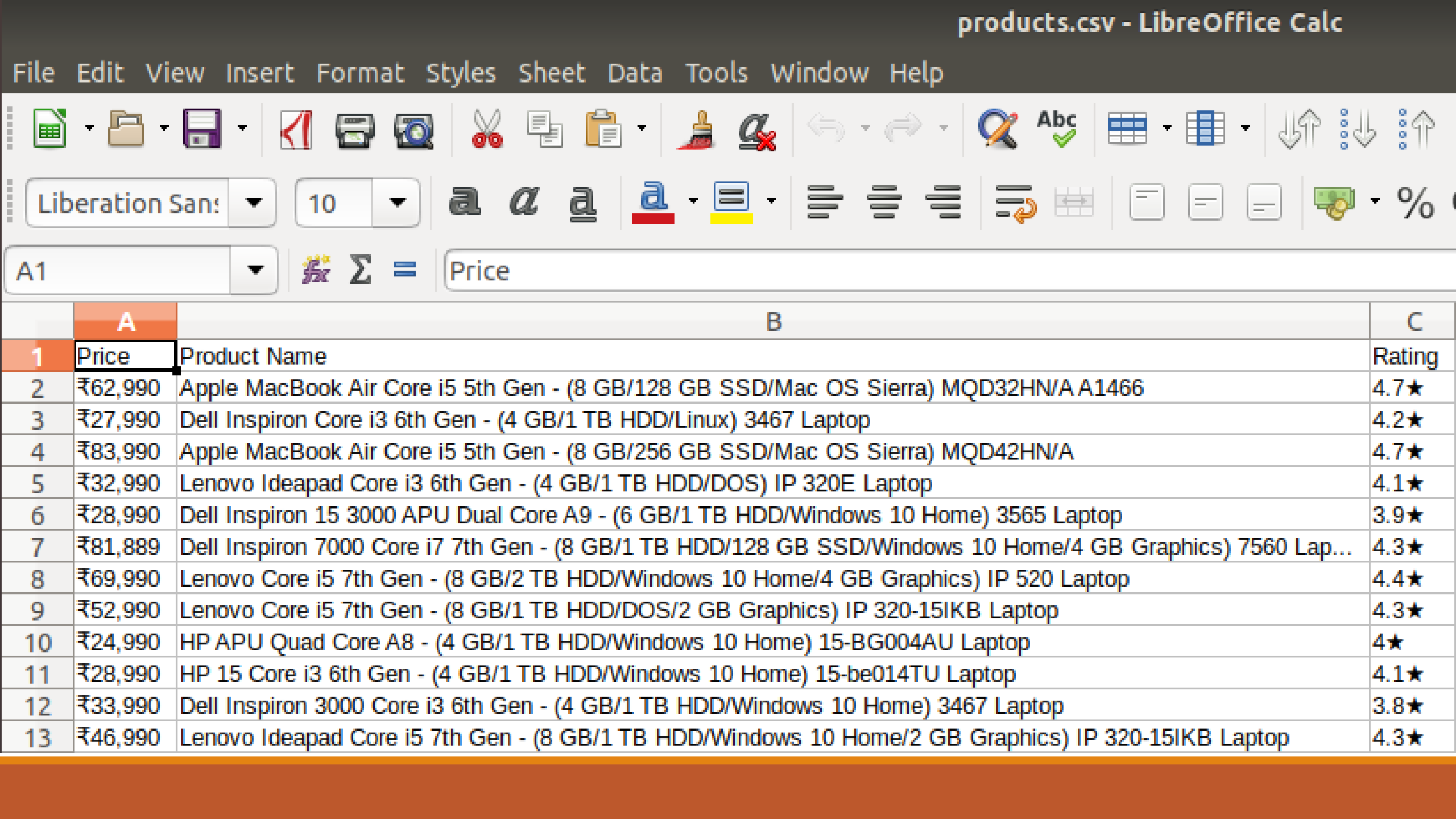
</h1>

<div id="qpeTitleTag_feature_div" class="celwidget" data-feature-name="qpeTitleTag" data-csa-c-id="ygvf
s3s-kajqr0" data-cel-widget="qpeTitleTag_feature_div">
Styles
Computed
Layout
Event Listeners
Filter
:hov .cls
element.style {
}
#productTitle { 61NWx6dKkUL…&VLwB86rJ:
font-size: 19px!important;
}
.product-title-word-break { ref=s9_acss…24031:114
word-break: break-word;
}
.a-size-large { 11EIQ5IGqaL…/AmazonUI:
text-rendering: optimizeLegibility;
}
.a-size-large { 11EIQ5IGqaL…/AmazonUI:
font-size: 24px!important;
line-height: 32px!important;
}
* { 11EIQ5IGqaL…AmazonUI:1
-moz-box-sizing: border-box;
-webkit-box-sizing: border-box;
box-sizing: border-box;
}
Inherited from h1#title.a-s…
h1 { 11EIQ5IGqaL…/AmazonUI:
font-weight: 400;
font-size: 28px;
line-height: 36px;
}
h1, h2, h3, h4 { 11EIQ5IGqaL…/AmazonUI:
text-rendering: optimizeLegibility;
}
h1 { user agent styleshee
font-size: 2em;
font-weight: bold;
}
div#titleSection.a-section.a-spacing-none
h1#title.a-size-large.a-spacing-none
span#productTitle.a-size-large.product-title-word-break
Activate Windows
Go to Settings to activate Windows.

Step 5: Run the code by using the below command:

python web-scrap.py

Step 6: A file called “products.csv” is created, which contains the extracted data.



	A	B	C
1	Price	Product Name	Rating
2	₹62,990	Apple MacBook Air Core i5 5th Gen - (8 GB/128 GB SSD/Mac OS Sierra) MQD32HN/A A1466	4.7★
3	₹27,990	Dell Inspiron Core i3 6th Gen - (4 GB/1 TB HDD/Linux) 3467 Laptop	4.2★
4	₹83,990	Apple MacBook Air Core i5 5th Gen - (8 GB/256 GB SSD/Mac OS Sierra) MQD42HN/A	4.7★
5	₹32,990	Lenovo Ideapad Core i3 6th Gen - (4 GB/1 TB HDD/DOS) IP 320E Laptop	4.1★
6	₹28,990	Dell Inspiron 15 3000 APU Dual Core A9 - (6 GB/1 TB HDD/Windows 10 Home) 3565 Laptop	3.9★
7	₹81,889	Dell Inspiron 7000 Core i7 7th Gen - (8 GB/1 TB HDD/128 GB SSD/Windows 10 Home/4 GB Graphics) 7560 Lap...	4.3★
8	₹69,990	Lenovo Core i5 7th Gen - (8 GB/2 TB HDD/Windows 10 Home/4 GB Graphics) IP 520 Laptop	4.4★
9	₹52,990	Lenovo Core i5 7th Gen - (8 GB/1 TB HDD/DOS/2 GB Graphics) IP 320-15IKB Laptop	4.3★
10	₹24,990	HP APU Quad Core A8 - (4 GB/1 TB HDD/Windows 10 Home) 15-BG004AU Laptop	4★
11	₹28,990	HP 15 Core i3 6th Gen - (4 GB/1 TB HDD/Windows 10 Home) 15-be014TU Laptop	4.1★
12	₹33,990	Dell Inspiron 3000 Core i3 6th Gen - (4 GB/1 TB HDD/Windows 10 Home) 3467 Laptop	3.8★
13	₹46,990	Lenovo Ideapad Core i5 7th Gen - (8 GB/1 TB HDD/Windows 10 Home/2 GB Graphics) IP 320-15IKB Laptop	4.3★