



UNIVERSITÀ
degli STUDI
di CATANIA

DIPARTIMENTO DI
MATEMATICA e INFORMATICA

Social Media Data Analysis 2023/2024

Web Scraping

Francesco Ragusa

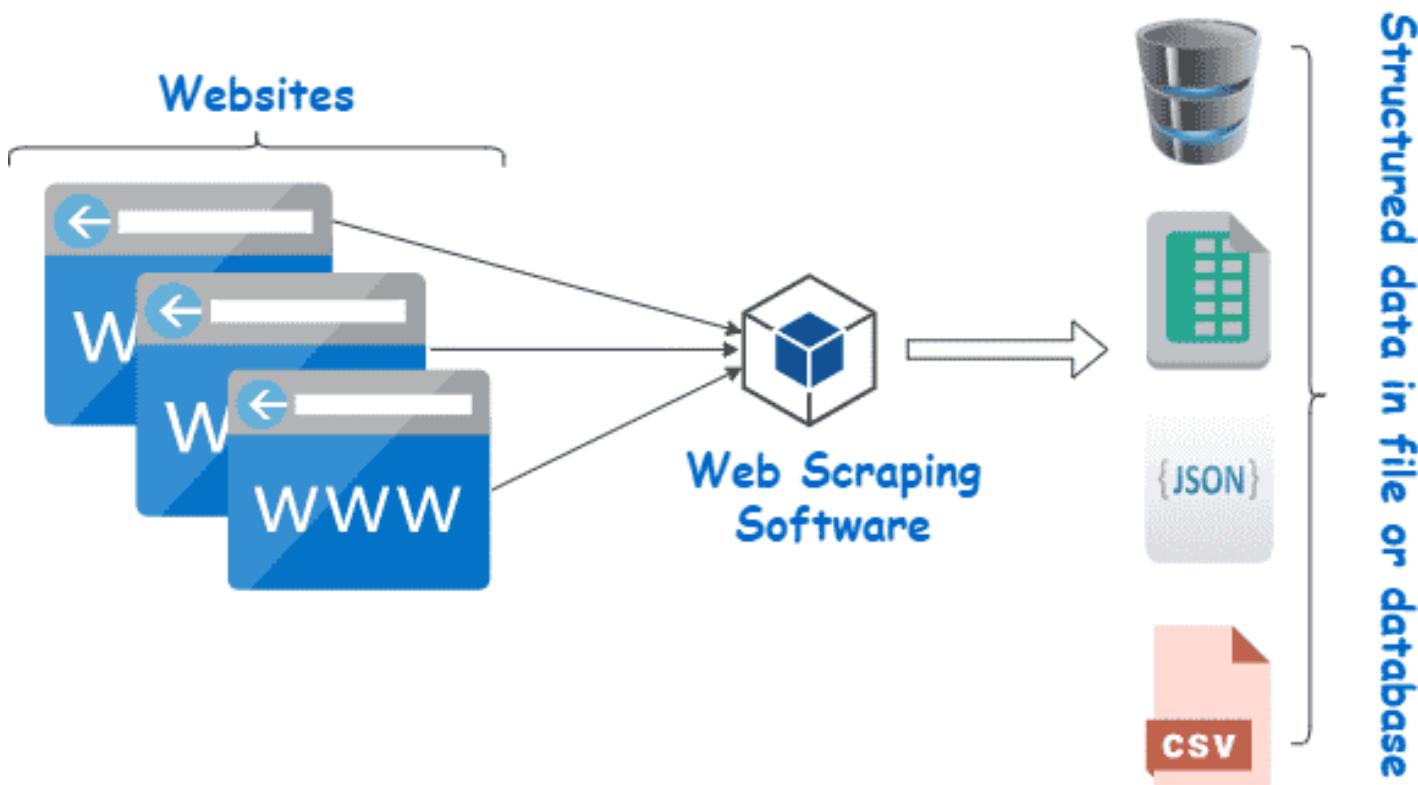
francesco.ragusa@unict.it

<https://iplab.dmi.unict.it/ragusa/>

<https://iplab.dmi.unict.it/fpv/>



Definition



Web scraping is the process to extract data from a website and export it into a structured format

Applications

- Price Comparison



Applications

- Price Comparison



- Email address gathering



Applications

- Price Comparison



- Email address gathering



- Social Media Scraping



Applications

- Price Comparison



- Email address gathering



- Social Media Scraping



- Research and Development



Applications

- Price Comparison



- Email address gathering



- Social Media Scraping



- Research and Development



- Job listings



Is it legal?

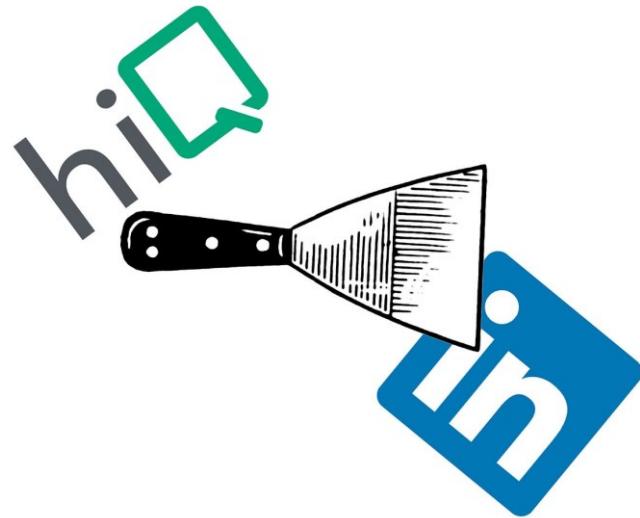
- How do you use the extracted data?
- Are you violating the ‘Terms & Conditions’ statements?
- Are you violating copyright of the extracted data?

Is it legal?

The data displayed by most of the websites are generally accessible to the public

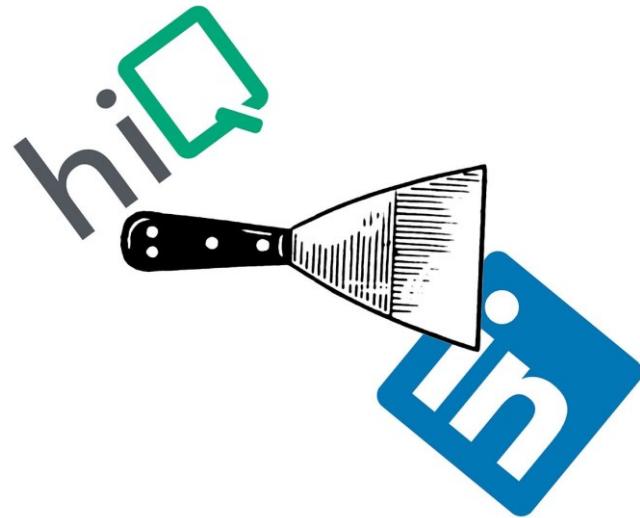
Is it legal?

The data displayed by most of the websites are generally accessible to the public



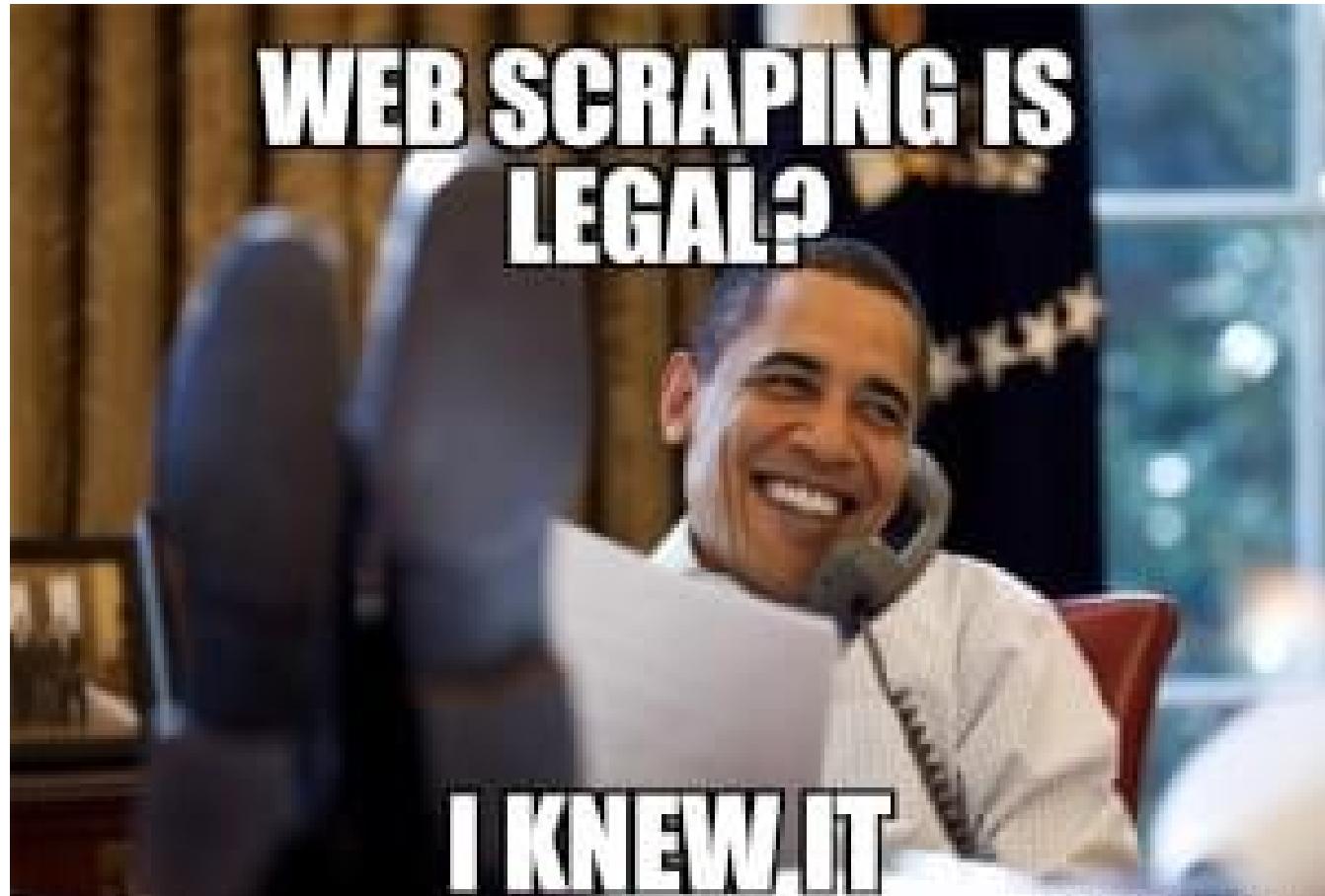
Is it legal?

The data displayed by most of the websites are generally accessible to the public



POWER VENTURES, INC.

Is it legal?



Is it legal?

Robots.txt

A text file created to instruct web robots (typically search engine robots) how to crawl pages on their website

These crawl instructions are specified by “disallowing” or “allowing” the behavior of certain (or all) user agents

User-agent: [user-agent name] Disallow: [URL string not to be crawled]

Robots.txt Syntax

- **User-agent:** The specific web crawler to which you're giving crawl instructions (usually a search engine). A list of most user agents can be found:
<http://www.robotstxt.org/db.html>
- **Disallow:** The command used to tell a user-agent not to crawl particular URL. Only one "Disallow:" line is allowed for each URL.
- **Allow** (Only applicable for Googlebot): The command to tell Googlebot it can access a page or subfolder even though its parent page or subfolder may be disallowed.
- **Crawl-delay:** How many seconds a crawler should wait before loading and crawling page content.
- **Sitemap:** Used to call out the location of any XML sitemap(s) associated with this URL.

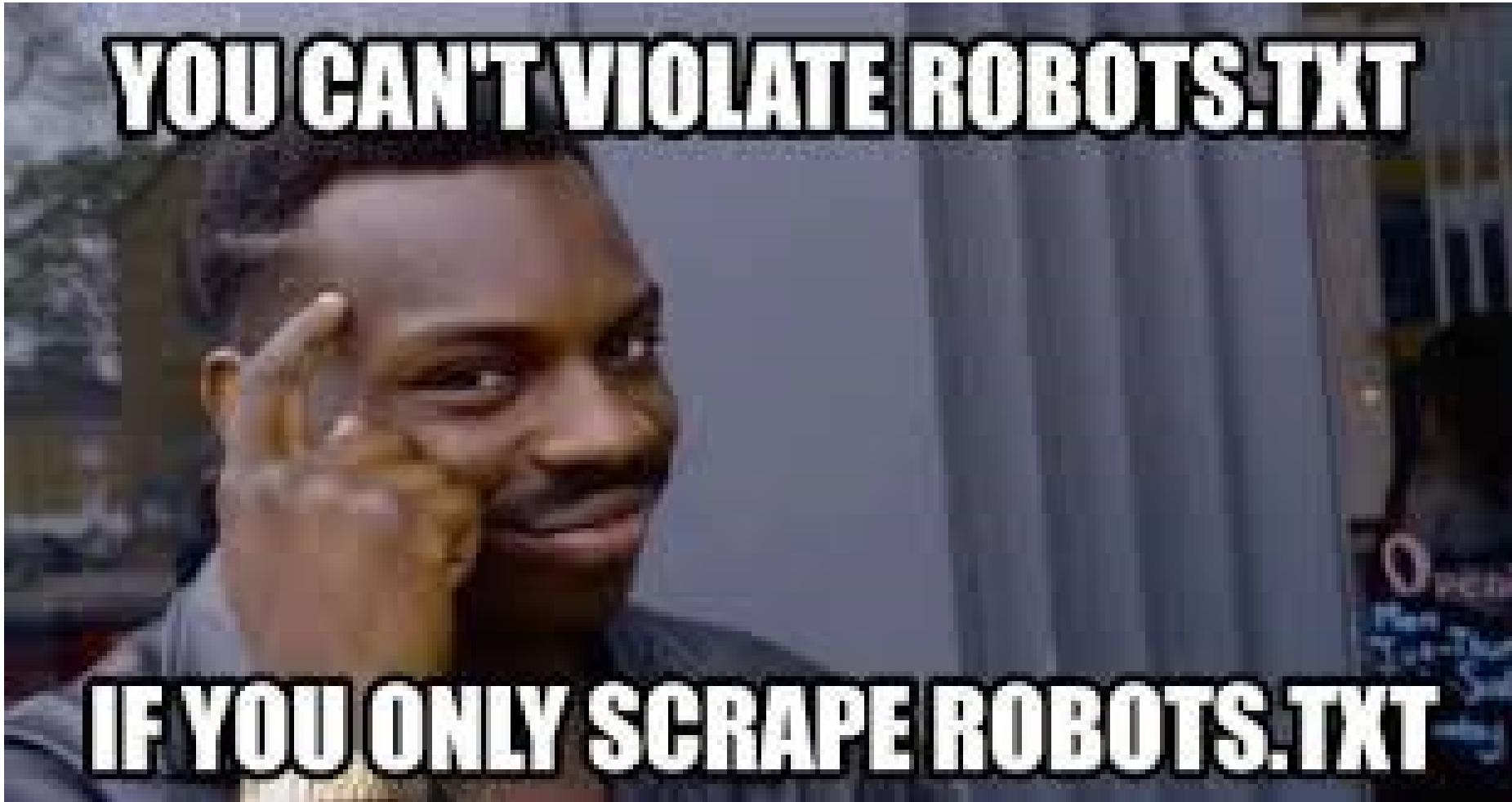
Robots.txt Syntax

```
# Example 1: Block only Googlebot
User-agent: Googlebot
Disallow: /

# Example 2: Block Googlebot and Adsbot
User-agent: Googlebot
User-agent: AdsBot-Google
Disallow: /

# Example 3: Block all crawlers except AdsBot (AdsBot crawlers must be named explicitly)
User-agent: *
Disallow: /
```

Robots.txt Syntax



DOM

The Document Object Model (DOM) is a ***programming interface*** for **HTML(HyperText Markup Language)** and **XML(Extensible markup language)** documents.

It defines the **logical structure** of documents and the way a document is accessed and manipulated.

This **model** allows JavaScript to access the text content and elements of the website **document as objects**.

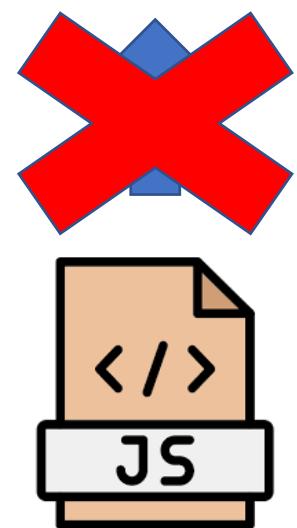
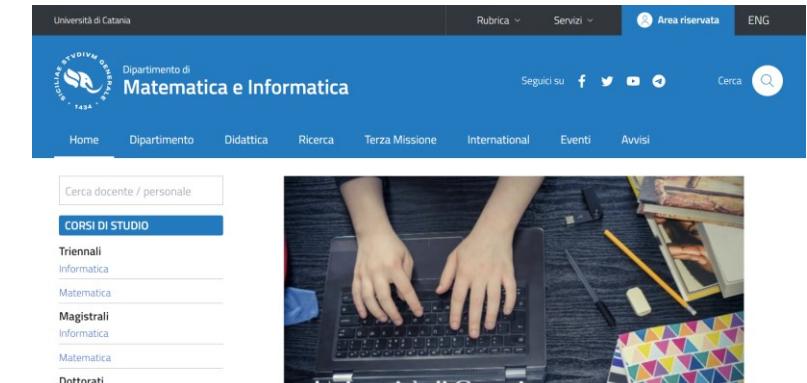
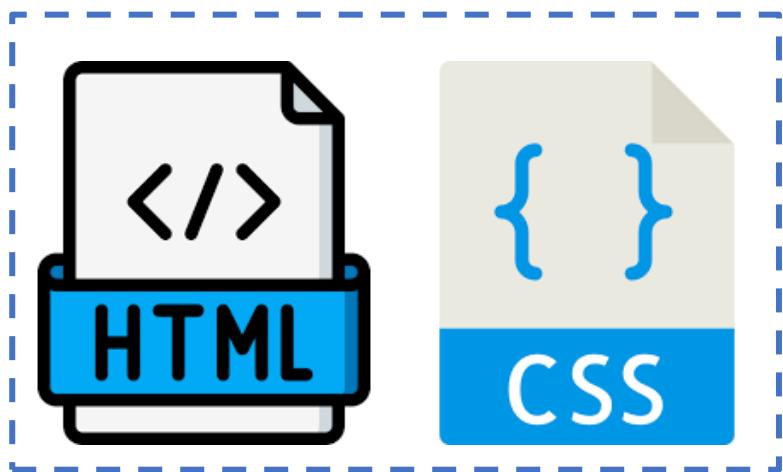
Document Object Model is an API that represents and interacts with HTML or XML documents.

DOM

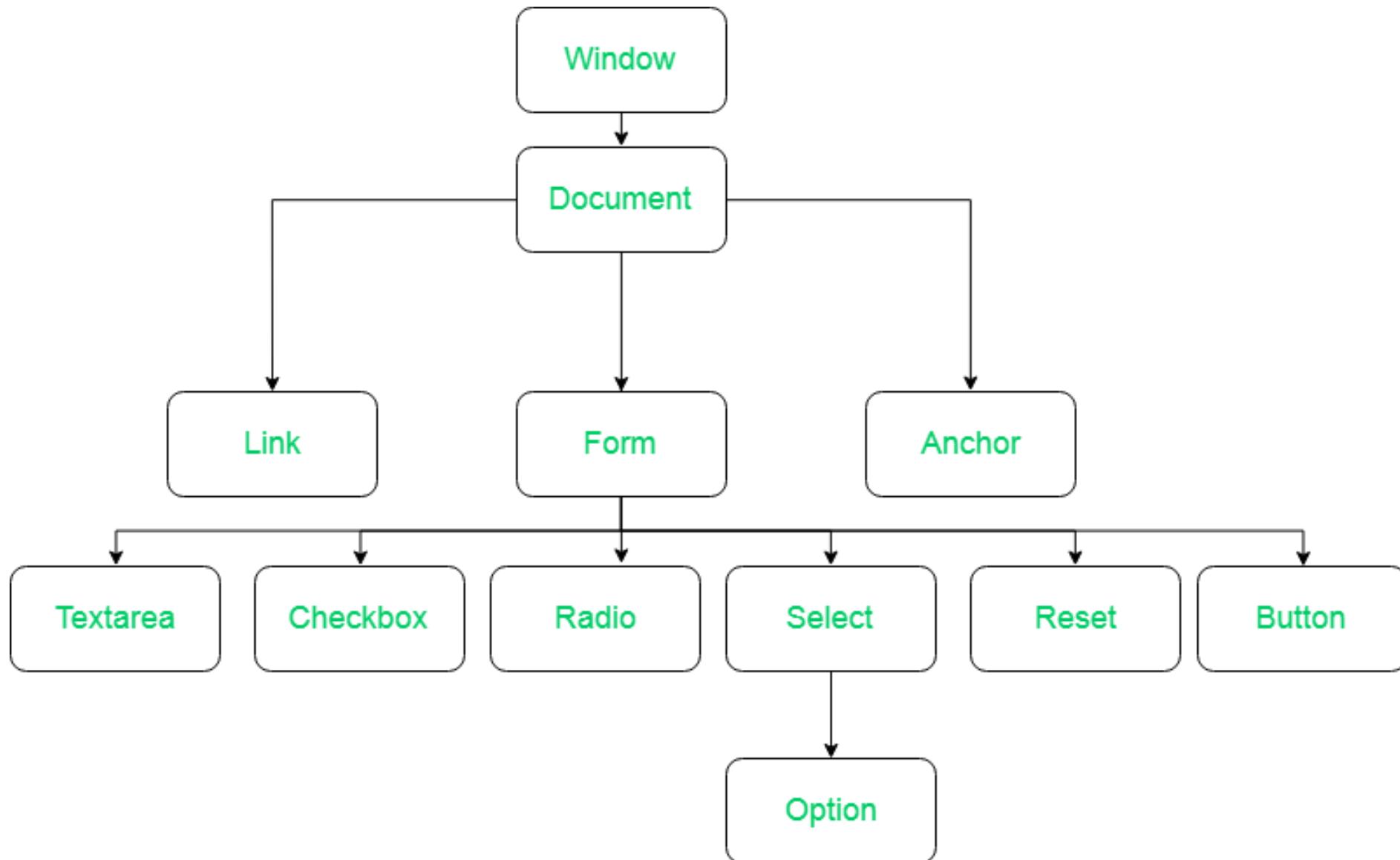
[Amazon.it: elettronica, libri, musica, fashion,
videogiochi, DVD e tanto altro](#)

DOM

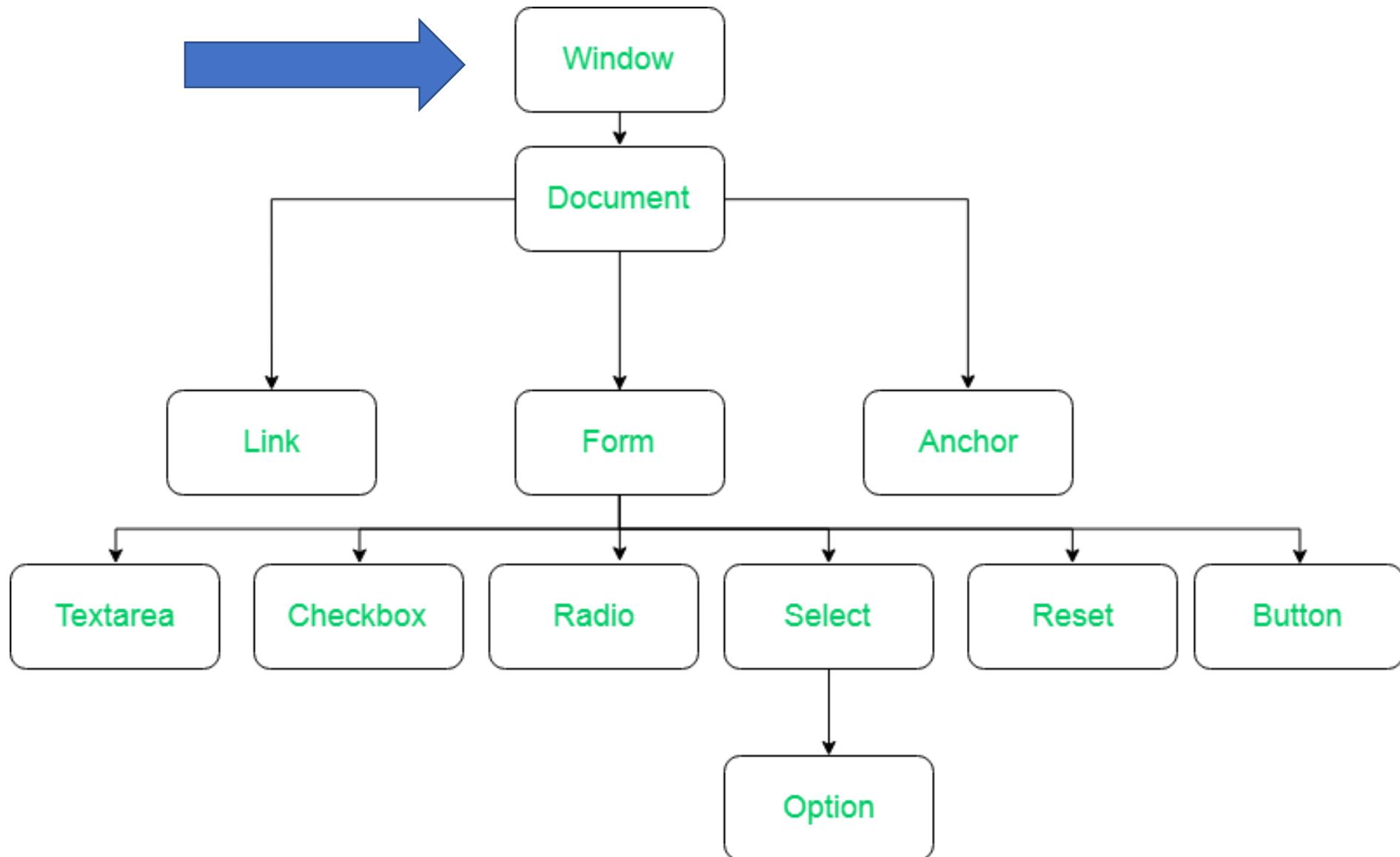
Web Site



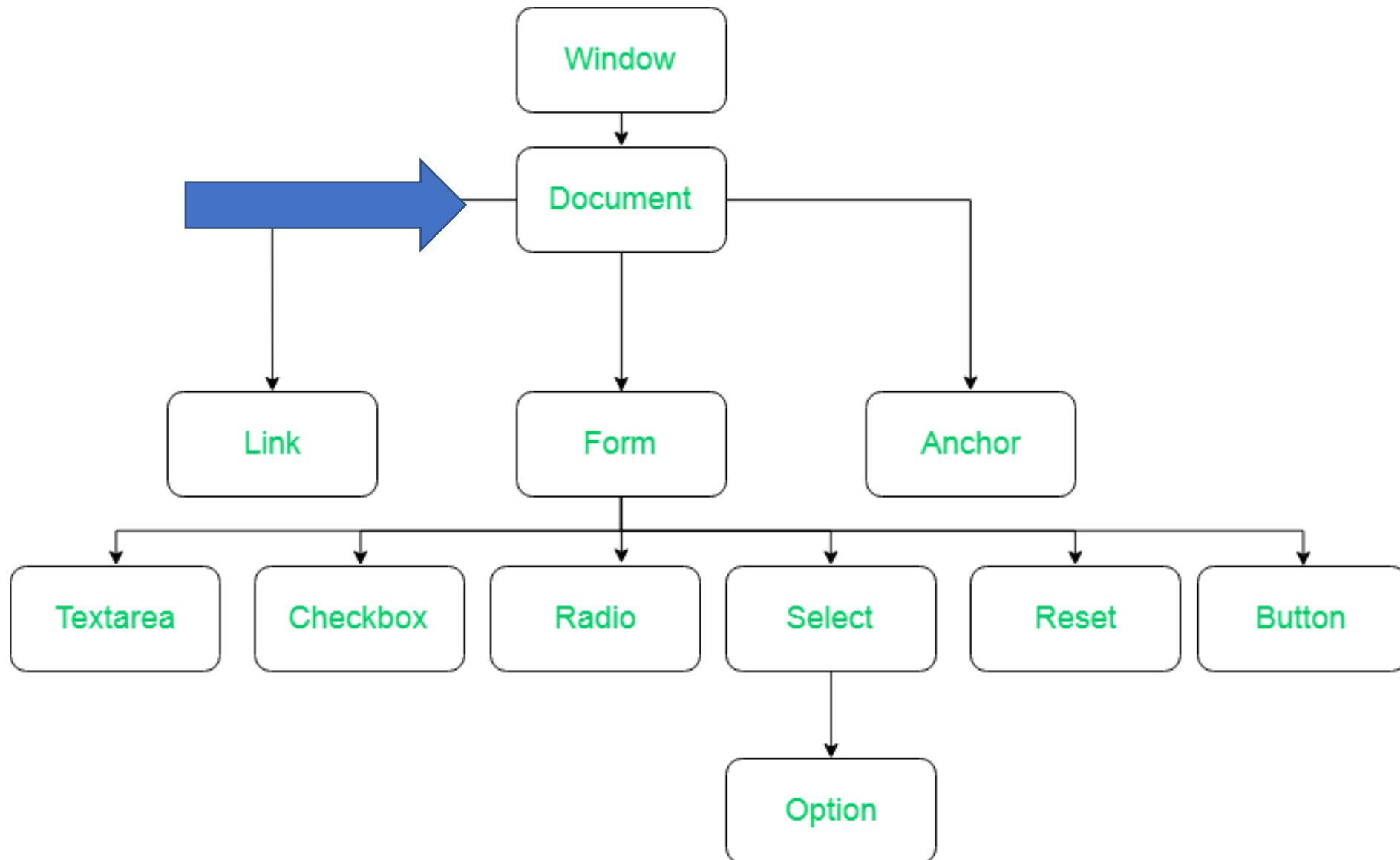
Properties



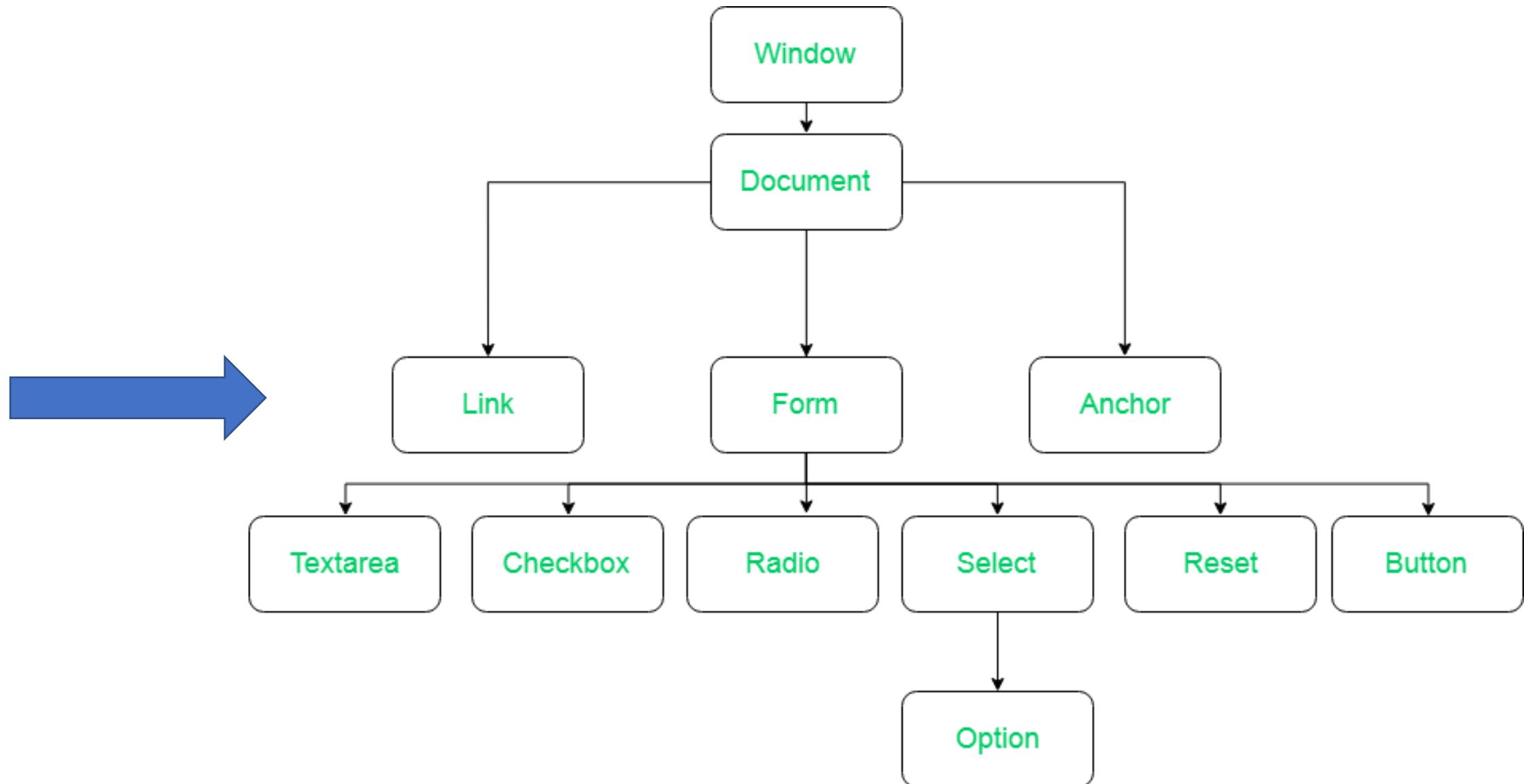
Properties



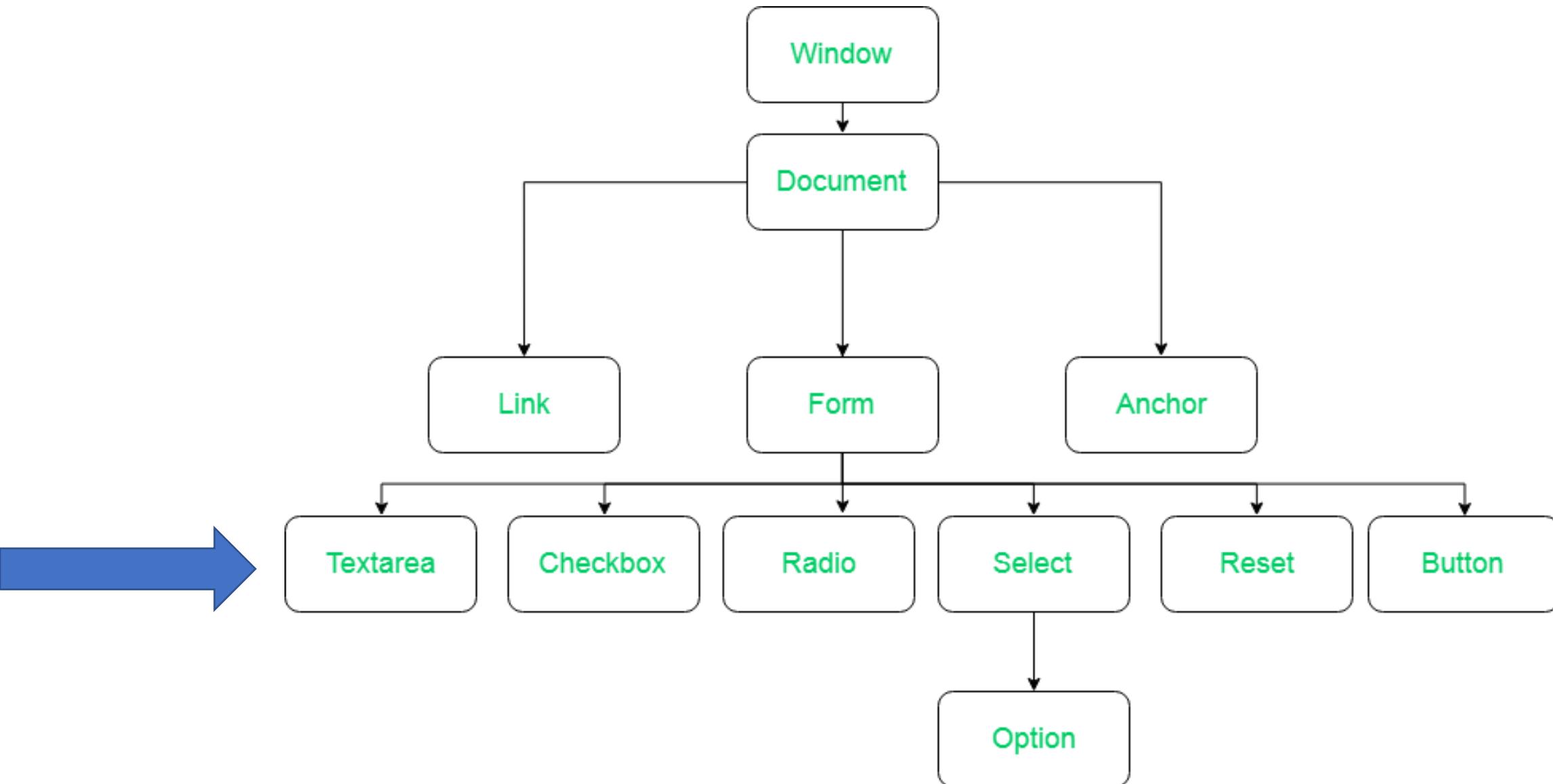
Properties



Properties



Properties



Methods

- **write**(“string”): Writes the given string on the document
- **getElementById()**: returns the element having the given id value
- **getElementsByName()**: returns all the elements having the given name value
- **getElementsByTagName()**: returns all the elements having the given tag name
- .
- **getElementsByClassName()**: returns all the elements having the given class name

Methods

SMDA Course

This is a lesson on Web Scraping.

This example illustrates the **getElementById** method.

The introduction is: This is a lesson on Web Scraping.

Methods

getElementById()

```
<!DOCTYPE html>
<html>

<body>
    <h2>SMDA Course</h2>

    <!-- Finding the HTML Elements by their Id in DOM -->
    <p id="intro">This is a lesson on Web Scraping.</p>
    <p>This example illustrates the <b>getElementById</b> method.</p>
    <p id="demo"></p>
    <script>
        const element = document.getElementById("intro");
        document.getElementById("demo").innerHTML =
            "The introduction is: " + element.innerHTML;
    </script>
</body>

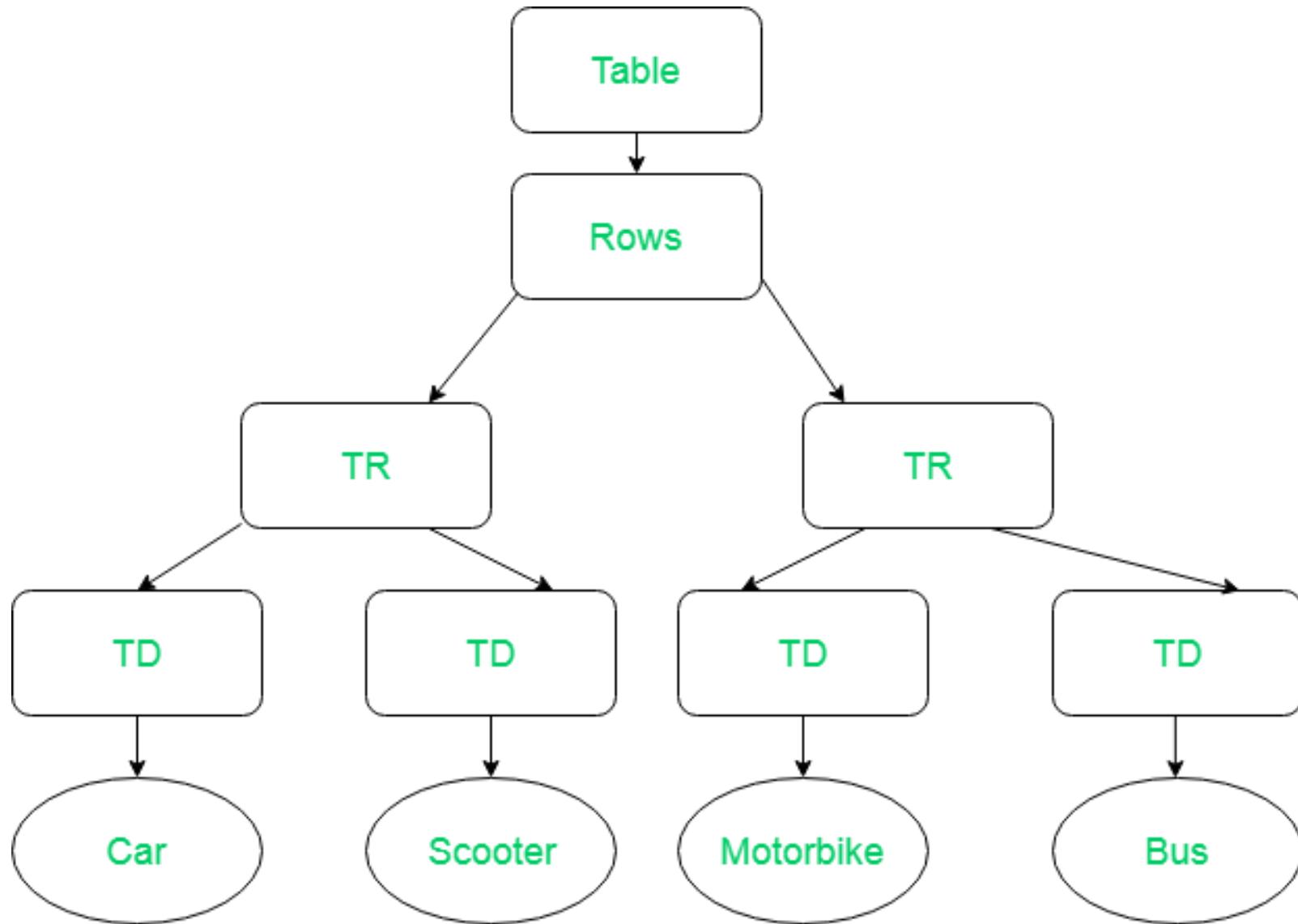
</html>
```

Tree Structure

Car	Scooter
MotorBike	Bus

```
<table>
  <ROWS>
    <tr>
      <td>Car</td>
      <td>Scooter</td>
    </tr>
    <tr>
      <td>MotorBike</td>
      <td>Bus</td>
    </tr>
  </ROWS>
</table>
```

Tree Structure



```
<!DOCTYPE html>
<html>

<head>
    <title>DOM manipulation</title>
</head>

<body>
    <label>Enter Value 1: </label>
    <input type="text" id="val1" />
    <br />
    <br />
    <label>Enter Value 2: </label>
    <input type=".text" id="val2" />
    <br />
    <button onclick="getAdd()">Click To Add</button>
    <p id="result"></p>
    <script type="text/javascript">
        function getAdd() {

            // Fetch the value of input with id val1
            const num1 = Number(document.getElementById("val1").value);

            // Fetch the value of input with id val2
            const num2 = Number(document.getElementById("val2").value);
            const add = num1 + num2;
            console.log(add);

            // Displays the result in paragraph using dom
            document.getElementById("result").innerHTML = "Addition : " + add;

            // Changes the color of paragraph tag with red
            document.getElementById("result").style.color = "red";
        }
    </script>
</body>

</html>
```

Enter Value 1: 25

Enter Value 2: 30

Click To Add

Addition : 55

What DOM is not?

The Document Object Model is not used to describe **objects in XML or HTML** whereas the DOM describes XML and HTML documents as objects.

The Document Object Model is not represented by a **set of data structures**; it is an interface that specifies object representation.

The Document Object Model does not show the **criticality of objects** in documents i.e it doesn't have information about which object in the document is appropriate to the context and which is not.

Web Scraping with Python



Webpages



Web Scraping



Structured Data

Why Python?



Ease of Use: Python programming is simple to code. You do not have to add semi-colons “;” or curly-braces “{}” anywhere.

Large Collection of Libraries: Python has a huge collection of libraries such as Numpy, Matplotlib, Pandas etc.

Duck typing: In Python, you don’t have to define datatypes for variables, you can directly use the variables wherever required.

Easily Understandable Syntax: Python syntax is easily understandable mainly because reading a Python code is very similar to reading a statement in English.

Small code, large task: Web scraping is used to save time.

Selenium



Selenium is an open-source automated testing framework used to validate web applications across different browsers and platforms.

It is an executable module that runs a script on a browser instance and hence is also called **headless browser scraping**.



Selenium with Python

Why Selenium?

Today selenium is mainly used for web scraping and automation purposes.

- clicking on buttons*
- filling forms*
- scrolling*
- taking a screenshot*

E-commerce Web Scraping

SAVE MORE ON APP SELL ON LAZADA CUSTOMER CARE TRACK MY ORDER LOGIN SIGNUP

Lazada

Search in Lazada

APPLY NOW WITH Lazada Loans

Electronic Devices
Electronic Accessories
TV & Home Appliances
Health & Beauty
Babies & Toys
Groceries & Pets
Home & Living
Women's Fashion & Access...
Men's Fashion & Accessories
Kid's Fashion & Accessories
Sports & Lifestyle
Automotive & Motorcycles

10.10 TIPID FEST OCT 10 - 14

100% FREE SHIPPING No Min. Spend

SHOP NOW >

Load & eStore LazMall Vouchers

Flash Sale

<https://www.lazada.com.ph/>

E-commerce Web Scraping

```
# Web Scraping
from selenium import webdriver
from selenium.common.exceptions import *
# Data manipulation
import pandas as pd
# Visualization
import matplotlib.pyplot as plt
import seaborn as sns
```

E-commerce Web Scraping

```
webdriver_path = 'C://Users//me//chromedriver.exe' # Enter the file directory of the Chromedriver
Lazada_url = 'https://www.lazada.com.ph'
search_item = 'Nescafe Gold refill 170g'
```

E-commerce Web Scraping

```
# Select custom Chrome options
options = webdriver.ChromeOptions()
options.add_argument('--headless')
options.add_argument('start-maximized')
options.add_argument('disable-infobars')
options.add_argument('--disable-extensions')
# Open the Chrome browser
browser = webdriver.Chrome(webdriver_path, options=options)
browser.get(Lazada_url)
```

E-commerce Web Scraping

But how do you identify which elements to find?

An easy way to do this is to use Chrome's very own inspect tool

E-commerce Web Scraping

The screenshot shows a web browser window for Lazada PH. The URL is lazada.com.ph. The page features a search bar at the top with the placeholder "Search in Lazada". Below the search bar is a large promotional banner for the "10.10 TIPID FEST OCT 10 - 14". The banner highlights "UP TO ₱1,000 LazBonus Voucher". On the left side, there is a sidebar with various product categories: Electronic Devices, Electronic Accessories, TV & Home Appliances, Health & Beauty, Babies & Toys, Groceries & Pets, Home & Living, Women's Fashion & Access..., Men's Fashion & Accessories, Kid's Fashion & Accessories, Sports & Lifestyle, and Automotive & Motorcycles. At the bottom of the page, there are links for "Load & eStore", "LazMall", "Flash Sale", and "Messages". The right side of the screen shows the browser's developer tools, specifically the "Elements" tab, which displays the HTML structure of the search bar area. The "Event Listeners" tab is also visible, showing 139 issues related to permissions policy.

Lazada PH: Shop and get up to ₱1,000 off on LazBonus Voucher during the 10.10 TIPID FEST from Oct 10 - 14. Search in Lazada.

SAVE MORE ON APP SELL ON LAZADA

Electronic Devices
Electronic Accessories
TV & Home Appliances
Health & Beauty
Babies & Toys
Groceries & Pets
Home & Living
Women's Fashion & Access...
Men's Fashion & Accessories
Kid's Fashion & Accessories
Sports & Lifestyle
Automotive & Motorcycles

Load & eStore LazMall Flash Sale Messages

Elements Console Sources Network Event Listeners DOM Breakpoints Properties Accessibility

In the future, Permissions Policy feature join-ad-interest-advertiser:1 group will not be enabled by default in cross-origin iframes or same-origin iframes nested in cross-origin iframes. Calling joinAdInterestGroup will be rejected with NotAllowedError if it is not explicitly enabled.

E-commerce Web Scraping

```
search_bar = browser.find_element_by_id('q')
search_bar.send_keys(search_item).submit()
```

E-commerce Web Scraping

Nescafe Gold refill 170g - Buy N X +

lazada.com.ph/catalog/?q=Nescafe+Gold+refill+170g&_keyori=ss&from=input&spm=a2o4l.home.search.go.239eca18rwcVNZ

SAVE MORE ON APP SELL ON LAZADA CUSTOMER CARE TRACK MY ORDER LOGIN SIGNUP

Lazada Nescafe Gold refill 170g   APPLY NOW WITH Lazada Loans

Categories  Load & eStore LazMall Vouchers

Category: Nescafe Gold refill 170g

3-in-1 Coffee

Ground Coffee

Brand: Nescafe Gold

Service & Promotion: Cashback Sale Cash On Delivery Free Shipping

Location: Local Metro Manila

Sort By: Best Match  View:  

4 items found for "Nescafe Gold refill 170g"

Product	Description	Price	Action
	NESCAFE Gold Intense Coffee 170g REFILL (Exp: Oct 18 2023) makes 8...	₱1,102	
	AJDL sellPQKV.PH NESCAFE Gold Intense Coffee 170g REFILL (Exp: O...	₱1,164.97	
	NESCAFE Gold Instant Coffee 170g	₱499	
	TOP bfvbfi NESCAFE Gold Intense Coffee 170g REFILL (Exp: Oct 18 2023)	₱1,104	

E-commerce Web Scraping

Nescafe Gold refill 170g - Buy N +

lazada.com.ph/catalog/?q=Nescafe%20Gold%20refill%20170g

SAVE MORE ON APP SELL ON LINE Elements Console Sources Network » 1 11 7 ⚙️ :

Lazada

Nescafe Gold refill 170g

submersible pump 3hp | quantumin plus | brosko pampalaki spray | viagara tablets pfizer

Load & eStore LazMall Vouchers

Categories ▾

Brand

Nescafe Gold

Service & Promotion

Cashback
 Sale
 Cash On Delivery
 Free Shipping

Location

Local
 Metro Manila
 Cavite
 Abra

Price


span.ooOxS 57.14 × 20 fee 170g makes 8...
₱1,102
32% Off


AJDL sellIPQKV.PH NESCAFE Gold Intense Coffee 170g REFILL (Exp: 0...
₱1,164.97
42% Off

Metro Manila

Metro Manila

Messages

```
</div>
</div>


</div>
<div class="RfADT">
NESCAFE Gold Intense Coffee 170g REFILL \(Exp: Oct 18 2023\) makes 85 cups</a>
</div>
<div class="aBrP0">
₱1,102</span> == \$0
</div>
<div class="WNoq3" style="flex">
<div class="6uN7R" style="flex">
...
</div>
</div>
<div class="17mcb" style="flex">
<div.Bm3ON>
<div.Ms6aG>
<div.qmXQo>
<div.buTCK>
<div.aBrP0>


Console Issues



Default levels 10 hidden



24 Issues: 1 7 17



workbox Router is responding to: https://laz-g-cdn.alicdn.com/lzmod/desktop-footer/6.0.210/?pc/index.css



workbox Using CacheFirst to respond to 'http://laz-g-cdn.alicdn.com/lzmod/desktop-footer/6.0.210/?pc/index.css'


```

E-commerce Web Scraping

```
item_titles = browser.find_elements_by_class_name('RfADt')
item_prices = browser.find_elements_by_class_name('oo0xS')

# Initialize empty lists
titles_list = []
prices_list = []
# Loop over the item_titles and item_prices
for title in item_titles:
    titles_list.append(title.text)
for price in item_prices:
    prices_list.append(price.text)
```

E-commerce Web Scraping

```
[ 'NESCAFE Gold Intense Coffee 170g REFILL (Exp: Oct 18 2023) makes 85 cups',
  'AJDL sellPQKV.PH NESCAFE Gold Intense Coffee 170g REFILL (Exp: Oct 18 2023) makes 85 cups',
  'TOP bfvbfi NESCAFE Gold Intense Coffee 170g REFILL (Exp: Oct 18 2023) makes 85 cups',
  'NESCAFE Gold Instant Coffee 170g']
```

```
[ '₱1,102', '₱1,164.97', '₱1,104', '₱499' ]
```

```
dfL = pd.DataFrame(zip(titles_list, prices_list),
columns=[ 'ItemName', 'Price'])
```

	ItemName	Price
0	NESCAFE Gold Intense Coffee 170g REFILL (Exp: ...	₱1,102
1	AJDL sellPQKV.PH NESCAFE Gold Intense Coffee 1...	₱1,164.97
2	TOP bfvbfi NESCAFE Gold Intense Coffee 170g RE...	₱1,104
3	NESCAFE Gold Instant Coffee 170g	₱499

E-commerce Web Scraping

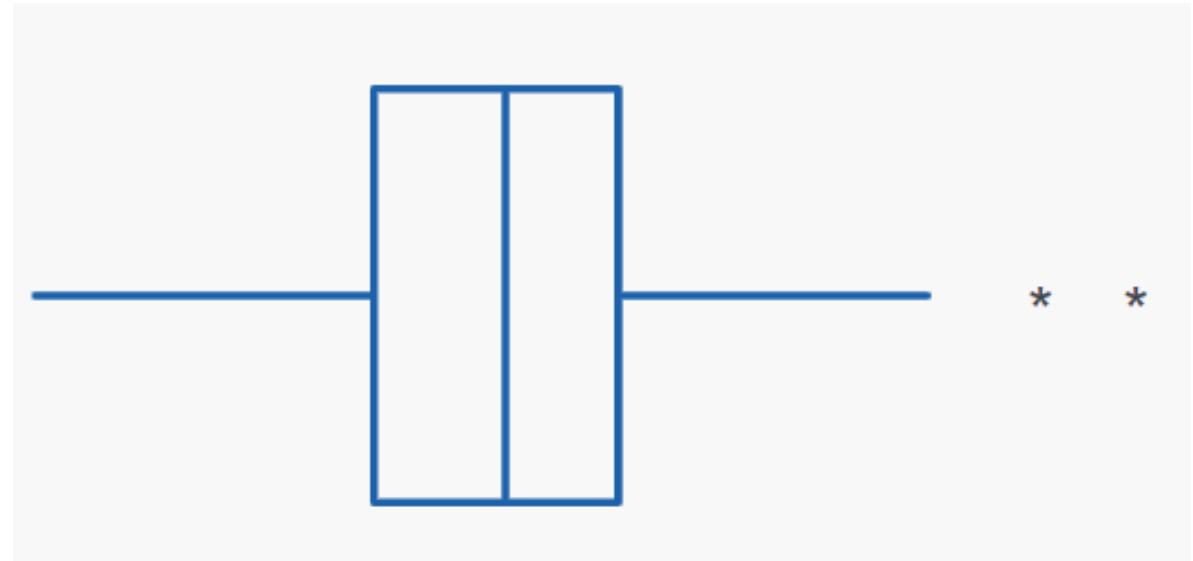
```
dfl['Price'] = dfl['Price'].str.replace('₱', '').astype(float)
```

```
dfl['Platform'] = 'Lazada'
```

		ItemName	Price	Platform
0	NESCAFE Gold Intense Coffee 170g REFILL (Exp: ...	1.102	Lazada	
1	AJDL sellPQKV.PH NESCAFE Gold Intense Coffee 1...	1.164	Lazada	
2	TOP bfvbfi NESCAFE Gold Intense Coffee 170g RE...	1.104	Lazada	
3	NESTLE NESCAFE Gold Instant Coffee 170g	0.499	Lazada	

E-commerce Web Scraping

- Lowest price
- Highest price
- Median price
- 25th and 75th percentile price

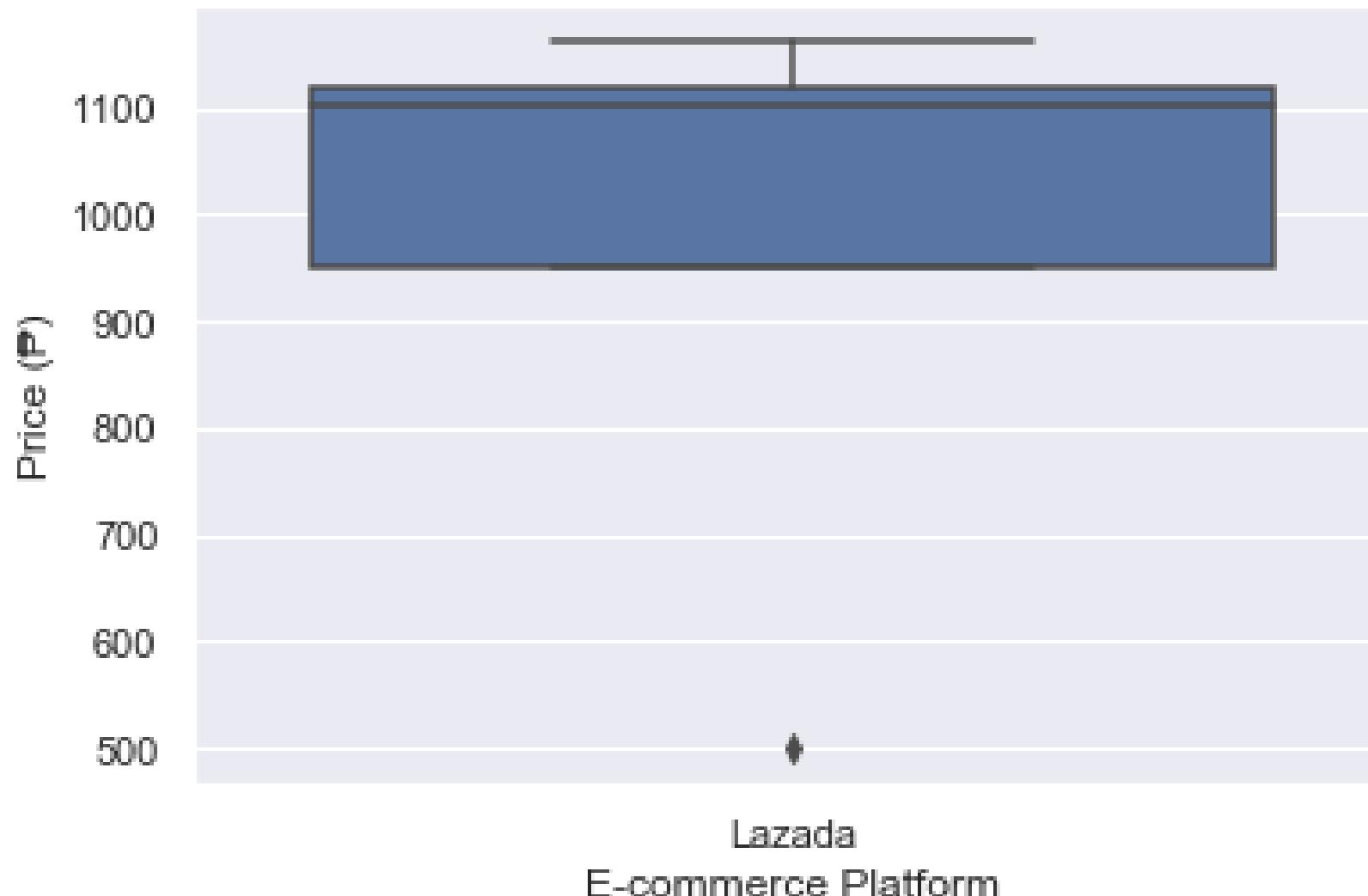


E-commerce Web Scraping

```
# Plot the chart
sns.set()
_ = sns.boxplot(x='Platform', y='Price', data=df1)
_ = plt.title('Comparison of Nescafe Gold Refill 170g prices between e-commerce platforms in Malaysia')
_ = plt.ylabel('Price (RM)')
_ = plt.xlabel('E-commerce Platform')
# Show the plot
plt.show()
```

E-commerce Web Scraping

Comparison of Nescafe Gold Refill 170g prices between e-commerce platforms in Malaysia



E-commerce Web Scraping

The screenshot shows the Shopee Philippines website (https://shopee.ph) with a focus on E-commerce Web Scraping. The page features a large orange header bar with the Shopee logo, a search bar, and a shopping cart icon. Navigation links include Seller Centre, Start Selling, Download, Notifications, Help, English, Sign Up, and Login. Promotional banners include "FREE SHIPPING ARAW-ARAW" (0 min. spend), "ShopeePay" (Over 200 fun games), and "UnionBank" (Up to ₱400 off). Below the header are category icons for Coins Rewards, Free Shipping, Shopee Mall, Vouchers, Shopee Beauty, Shopee Styles, Shopee Supermarket, Gadget Zone, Change of Mind Returns, and a "NEW ARRIVALS" section.

Seller Centre | Start Selling | Download | Follow us on [f](#) [o](#)

Notifications [Help](#) English [Sign Up](#) [Login](#)

Shopee Sign up and get 100% off on your first order

1 Peso Shoes 1 Peso Case 1 Peso Item Free Shipping Shoes For Basketball 1 Peso Sale Vepe Smoke Original Airpods Buy 1 Take 1

FREE SHIPPING ARAW-ARAW
₱0 MIN. SPEND

ShopeePay
Over 200 incredibly fun games, ad-free.
Use ShopeePay for your Apple Arcade subscription.

UnionBank
UP TO ₱400 OFF
USE CODE SHOPEEWED
EVERY WEDNESDAY

Coins Rewards Free Shipping Shopee Mall Vouchers Shopee Beauty Shopee Styles Shopee Supermarket Gadget Zone Change of Mind Returns

WIN GADGET OR 1,000 COINS!

NEW ARRIVALS

Popup Alerts



```
WebDriverWait(browser, 20).until(EC.element_to_be_clickable(  
    (By.XPATH, "//div[@class='shopee-modal__container']//button[text()='English']"))).click()
```

Multiple Prices

shopee.ph/search?keyword=nescafe%20gold%20refilll%20170g

Start Selling | Download | Follow us on

opee nescafe gold refilll 170g

1 Peso Shoes 1 Peso Case 1 Peso Item Free Shipping Shoes For Basketball 1 Peso Sale Vepe Smoke Ori

FILTER

Search result for 'nescafe gold refilll 170g'

Sort by **Relevance** Latest Top Sales Price

)

liances (1)

★

☆ & Up

☆ & Up

☆ & Up

☆ & Up


170g 1 Pack **GOLD**
170.67 x 19.2 mm
100% NESCAFE Gold Intense coffee 170g REFILL (Exp: Oct 2023)
₱3,122 - ₱4,508


170g 1 Pack **GOLD**
170.67 x 19.2 mm
100% NESCAFE Gold Intense coffee 170g REFILL (Exp: Oct 2023)
₱1,071

LIMITED STOCK


170g / 2 Packs **GOLD**
170g(twin pack) UGPA
₱1,923

Binondo, Metro Manila

Paranaque City, Metro Manila

Taguig City, Metro Manila

Elements Console Sources Network > x 1 ! 11 x 2 ⚙️ X

> <div style="pointer-events: none;"></div>

> <div class="KMyn8J" flex>

> <div class="dpiR4u" data-sqe="name"></div> flex

> <div class="hpDKMN" flex>

> ... <div class="vioxDd rVLWG6" = \$0

>

> ₱

> 3,122

>

> ₱

> 4,508

</div>

</div>

> <div class="ZnrnMl" flex>

> <div class="zGGwiV" flex>

</div>

_item a div.tWpFe2 div.VTjd7p.whlxGK div.KMyn8J div.hpDKMN div.vioxDd.rVLWG6

Styles Computed Layout Event Listeners DOM Breakpoints Properties Accessibility

Filter :hov .cls + ↻

Console What's New Issues

Highlights from the Chrome 118 update

Enhanced search

View all matches in a line of a minified file in search results and enjoy boosted search speed.

```
product_prices = browser.find_elements_by_xpath('//a/div/div[2]/div[2]/div[@*]/span[2]')
```

Item Name are not selectable

```
import requests
Shopee_url = 'https://shopee.com.my'
keyword_search = 'Nescafe Gold refill 170g'
headers = {
    'User-Agent': 'Chrome',
    'Referer': '{}search?keyword={}'.format(Shopee_url, keyword_search)
}
url = 'https://shopee.com.my/api/v2/search_items/?by=relevancy&keyword={}&limit=100&newest=0&order=desc&page_type=search'.format(keyword_search)
# Shopee API request
r = requests.get(url, headers = headers).json()
# Shopee scraping script
titles_list = []
prices_list = []
for item in r['items']:
    titles_list.append(item['name'])
    prices_list.append(item['price_min'])
```

E-commerce Web Scraping

```
dfS = pd.DataFrame(zip(titles_list, prices_list), columns=['ItemName', 'Price'])
```

	ItemName	Price
0	NESCAFE Gold REFILL (170g)	₱832
1	Nescafe Gold Refill 170g	₱732
2	Nescafe gold refill 170g Original Import	₱1191
3	NESCAFE Gold Intense Coffee 170g REFILL	₱1071
4	NESCAFE Gold Intense Coffee 170g REFILL	₱1082
5	NESCAFE Gold Intense Coffee 170g REFILL	₱1082
6	[Hot Sale] NESCAFE Gold Intense Coffee 170g RE...	₱1081

E-commerce Web Scraping

```
dfS['Platform'] = 'Shopee'  
# Concatenate the Dataframes  
df = pd.concat([dfL,dfS])
```

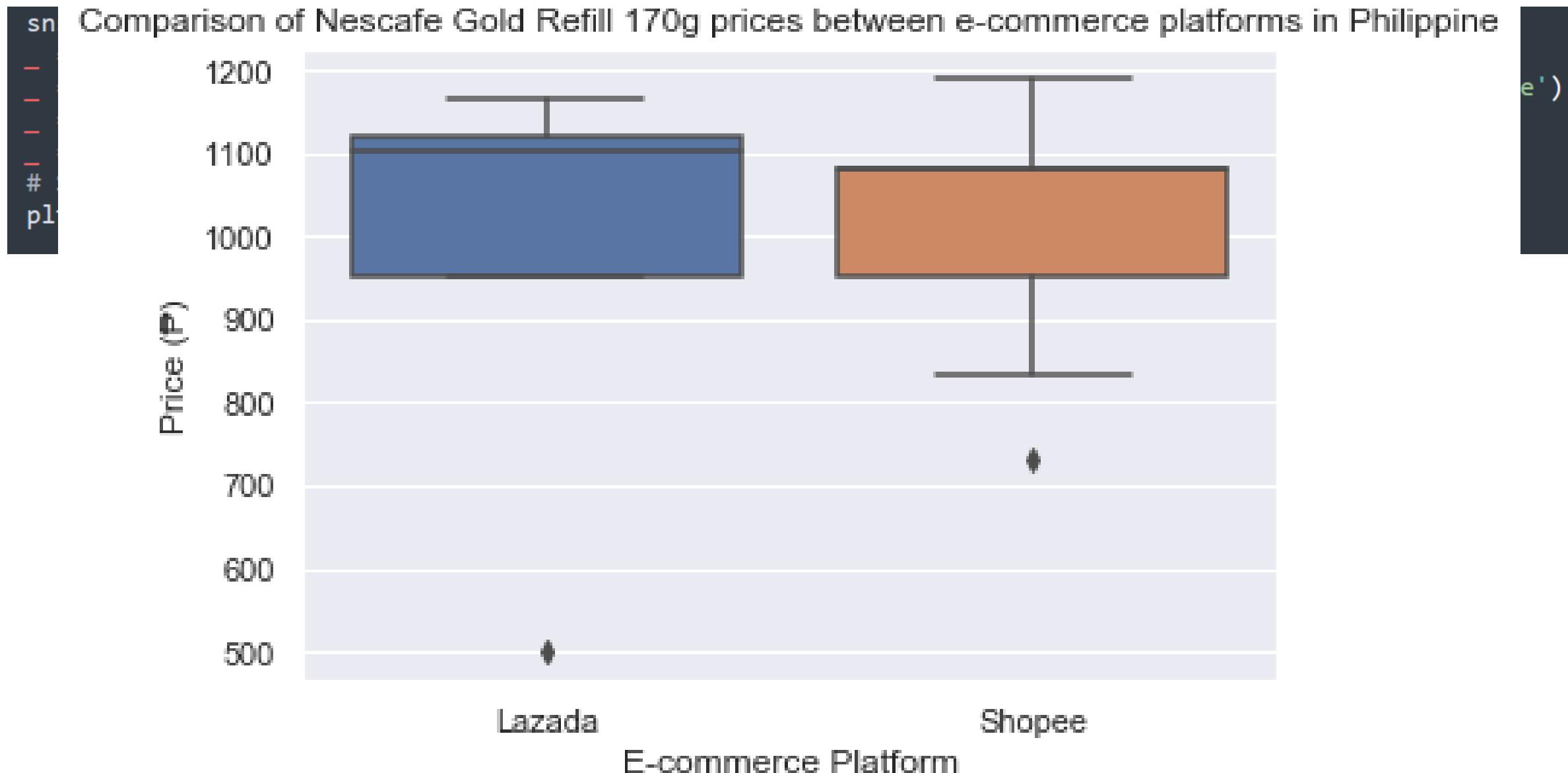
	ItemName	Price	Platform
0	NESCAFE Gold Intense Coffee 170g REFILL (Exp: ...	1102.0	Lazada
1	AJDL sellIPQKV.PH NESCAFE Gold Intense Coffee 1...	1164.0	Lazada
2	TOP bfvbfi NESCAFE Gold Intense Coffee 170g RE...	1104.0	Lazada
3	NESCAFE Gold Instant Coffee 170g	499.0	Lazada
0	NESCAFE Gold REFILL (170g)	832.0	Shopee
1	Nescafe Gold Refill 170g	732.0	Shopee
2	Nescafe gold refill 170g Original Import	1191.0	Shopee
3	NESCAFE Gold Intense Coffee 170g REFILL	1071.0	Shopee
4	NESCAFE Gold Intense Coffee 170g REFILL	1082.0	Shopee
5	NESCAFE Gold Intense Coffee 170g REFILL	1082.0	Shopee
6	[Hot Sale] NESCAFE Gold Intense Coffee 170g RE...	1081.0	Shopee

E-commerce Web Scraping

```
print(df.groupby(['Platform']).describe())
```

Platform	Price								
	count	mean	std	min	25%	50%	75%	max	
Lazada	4.0	967.250000	313.489367	499.0	951.25	1103.0	1119.0	1164.0	
Shopee	7.0	1010.142857	163.737013	732.0	951.50	1081.0	1082.0	1191.0	

E-commerce Web Scraping



Libraries



It allows us to make an http request to different websites

It opens a socket to the target website and asks them for their permission to connect

Beautiful Soup

[Beautiful Soup 4.4.0 documentation »](#)

[index](#)

Table Of Contents

Beautiful Soup Documentation

- [Getting help](#)

Quick Start

Installing Beautiful Soup

- [Problems after installation](#)
- [Installing a parser](#)

Making the soup

Kinds of objects

- [Tag](#)
 - [Name](#)
 - [Attributes](#)
 - [Multi-valued attributes](#)

[NavigableString](#)

[BeautifulSoup](#)

- [Comments and other special strings](#)

Navigating the tree

- [Going down](#)
 - [Navigating using tag names](#)
 - [.contents and .children](#)
 - [.descendants](#)
 - [.string](#)
 - [.strings and .strings_as_list](#)

Beautiful Soup Documentation

Beautiful Soup is a Python library for pulling data out of HTML and XML files. It works with your favorite parser to provide idiomatic ways of navigating, searching, and modifying the parse tree. It commonly saves programmers hours or days of work.

These instructions illustrate all major features of Beautiful Soup 4, with examples. I show you what the library is good for, how it works, how to use it, how to make it do what you want, and what to do when it violates your expectations.

The examples in this documentation should work the same way in Python 2.7 and Python 3.2.

You might be looking for the documentation for Beautiful Soup 3. If so, you should know that Beautiful Soup 3 is no longer being developed, and that Beautiful Soup 4 is recommended for all new projects. If you want to learn about the differences between Beautiful Soup 3 and Beautiful Soup 4, see [Porting code to BS4](#).

This documentation has been translated into other languages by Beautiful Soup users:

- 这篇文档当然还有中文版.
- このページは日本語で利用できます(外部リンク)
- 이 문서는 한국어 번역도 가능합니다. (외부 링크)

Getting help



Beautiful Soup Example

```
uClient=uRequest("https://www.thomann.de/it/ukulele_soprani.html")
page_html = uClient.read()
page_soup = soup(page_html)
containers=page_soup.findAll('div',{'class':'extensible-article'})
print(len(containers))
print(containers[0])
```

25

```
<div class="extensible-article list-view compare parent" rel="320760"> <div clas
s="left"> <div class="product-image image-block"> <a class="article-link link" h
ref="https://www.thomann.de/it/harley_benton_ukulele_uk_11dw_brown.htm"> 
```

Beautiful Soup Example

```
print(containers[0].findAll('span',{'class':'manufacturer'})[0].text)
print(containers[0].findAll('span',{'class':'model'})[0].text)

print()

print(containers[1].findAll('span',{'class':'manufacturer'})[0].text)
print(containers[1].findAll('span',{'class':'model'})[0].text)
```

Harley Benton

Ukulele UK-11DW Brown

Baton Rouge

V2-S sun

Beautiful Soup Example

The screenshot shows a web page from a music store. On the left, there's a sidebar with a "Su di noi" section featuring a photo of a man playing a guitar in a hallway, and text about the company's history (since 1954), family business, and client base. Below that is a "Servizio Made in Germany" section. A "La tua opinione" section contains a testimonial and a timestamp (D.S., 17.03.2018). The main content area displays four ukuleles:

- Steinberg UR22 MK2**: €115 / €159. Includes a Sengon neck, Ebonized fretboard, and immediate availability.
- Thomann Soprano Ukulele Standard**: €99. 4.5 stars. Includes Solid acacia top, sides, and neck/fretboard.
- Thomann Soprano Ukulele De Luxe**: €129. 4.5 stars. Includes De Luxe version, Solid acacia top, sides, and neck/fretboard.
- Lanikai Mahogany Soprano Ukulele MAHS**: €99. 4.5 stars. Includes Mahogany body, neck, and bridge, and Walnut fretboard and bridge.
- Flight TUS50 Travel Ukulele Walnut**: €49. 4.5 stars. Includes Durable and travel-friendly design, Walnut top, and Arched body.

At the bottom, there are buttons for "MOSTRA DI PIÙ" and "25".

The screenshot shows the browser's developer tools with the "Elements" tab selected. The DOM tree highlights a "div.page" element under "#resultPageNavigation". The element has a bounding box of 30.02 x 40 pixels. The "Styles" tab shows the CSS rules applied to this element, including "margin: 4px", "border: none", and "padding: 4px". The "Properties" tab shows the computed values for width (30.02) and height (40). The "Box Sizing" panel on the right shows the detailed box model with outer (orange), inner (green), and padding (blue) dimensions.

```
.rs-pagination-skin-boxes .container>.page, .rs-pagination-skin-boxes .container>.separator, .rs-pagination-skin-boxes .container>.page-separator, .rs-pagination-skin-boxes .container>.button { margin: 4px; border: none; padding: 4px; }
```

Addditional References

- [Web Scraping With Python - Full Guide to Python Web Scraping \(edureka.co\)](#)
- [What Is A Robots.txt File? Best Practices For Robot.txt Syntax – Moz](#)
- [DOM \(Document Object Model\) - GeeksforGeeks](#)