



Web Scraping

(IT069IU)

Le Duy Tan, Ph.D.

 ldtan@hcmiu.edu.vn

 leduytanit.com

Previously,

- Design Patterns:
 - Creational
 - Singleton
 - Factory Method
 - Structural
 - Behavioral



**Java programmer
says:**



**No problem, I can use
these seven design
patterns for that.**



Agenda

- **What is Web Scraping?**
- **Why is it important?**
- **How to use it safely and legally?**
- HTML, CSS
- Introduction to **JSoup Library**
- JSoup workflow
- VNExpress Scrapping Demo
- Manga Scrapping Demo
- Potential issues with Web Scraping

What is Web Scraping?



Web Pages



Web Scraping



Structured Data

Why is Web Scraping Used?

- Search engines
- Social Media Scraping
- Price intelligence
- Lead generation
- Research and Development
- Job listings
- Machine Learning
- Personal Hobby
 - Download music, manga, movies...

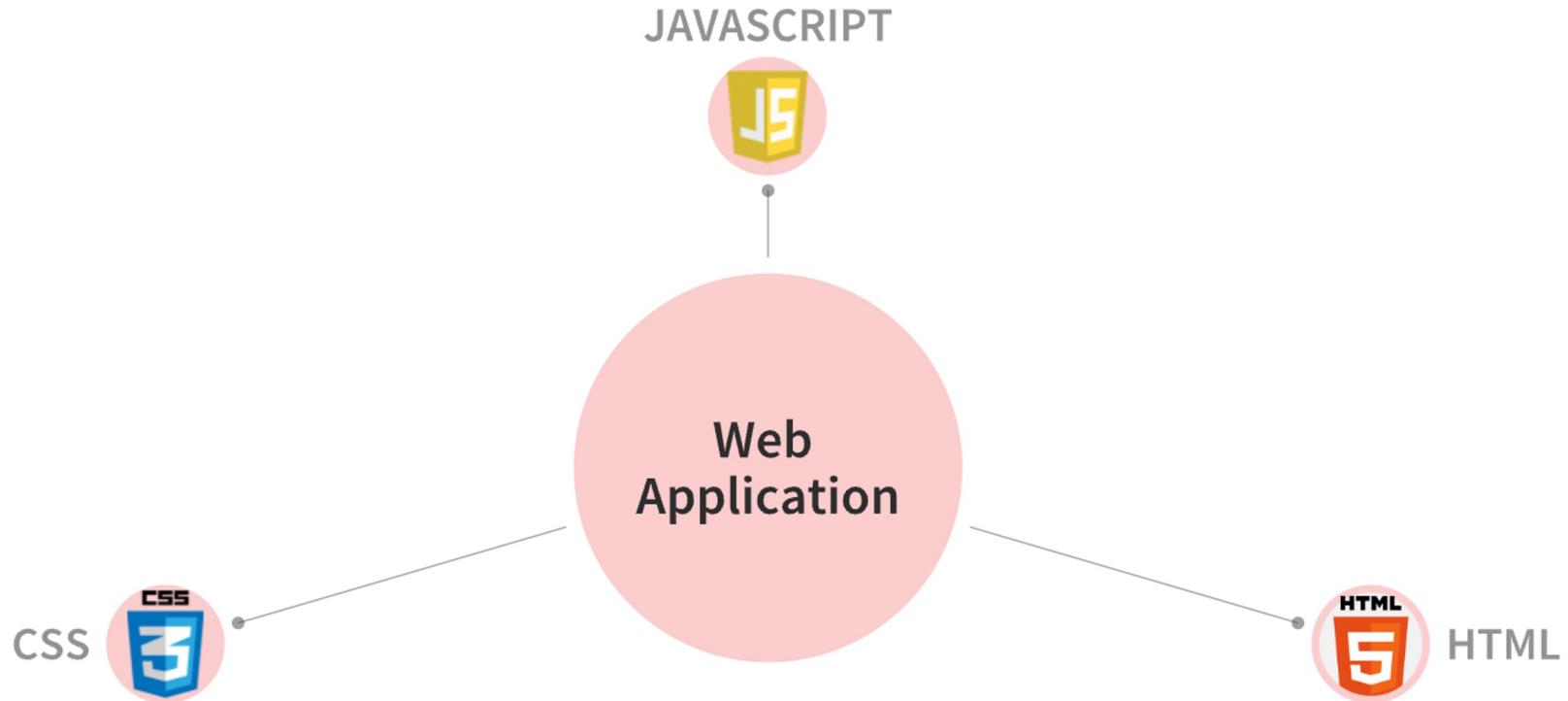
? Question: why don't we use official API from the websites instead of web scraping?



Website Components



Web Component Overview



Analogy of HTML, CSS & Javascript



HTML



CSS

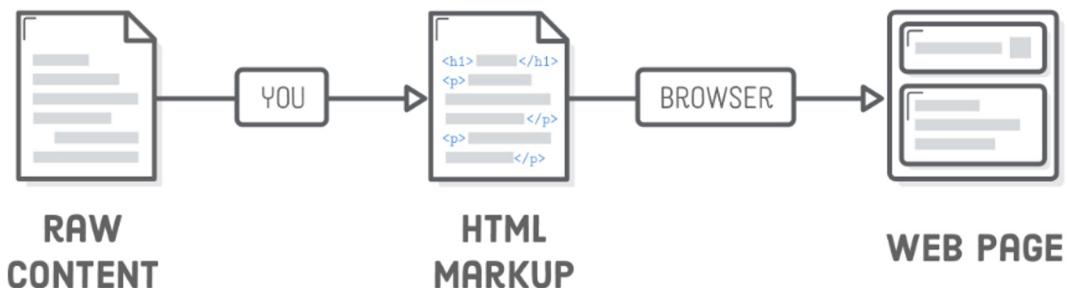
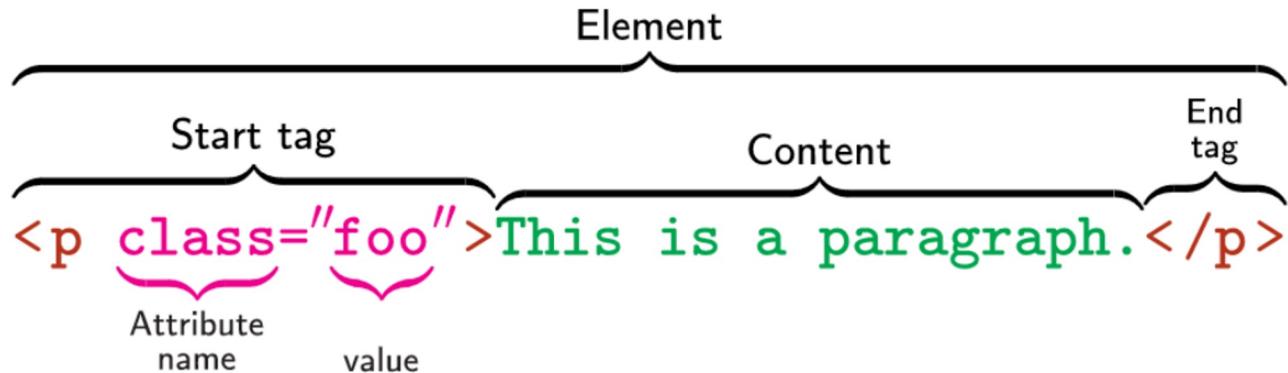


Javascript

What is HTML (Hyper Text Markup Language)?

- Hyper Text Markup Language (HTML)
- Every single web pages uses HTML
- HTML describes the page's content
- HTML is made up of elements, represented by tags

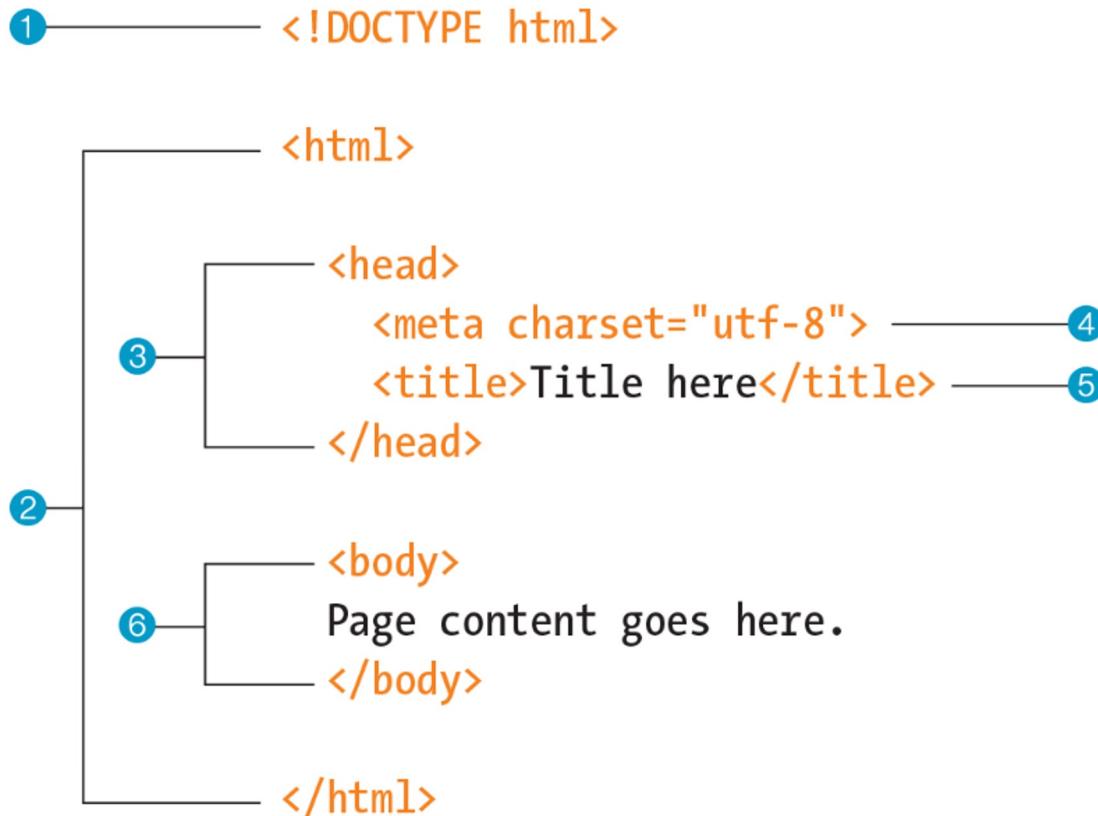
HTML Element



Common HTML Tags

- <head>
- <html>
- <body>
- <title>
- <div>
- <a>
- <link>
-
- <p>
-
- , ,
-
, <hr>
- <style>
- <form>, <input>
- <h1>, <h2>, <h3>
- <script>
- <table>, <tbody>, <tr>, <td>

Basic HTML Structure



Sample Basic HTML

A screenshot of a web browser window showing an HTML file named "lesson2.html". The browser interface includes a toolbar at the top with back, forward, and search buttons, and a status bar indicating the file path. The main area displays both the rendered HTML content and the source code. The rendered content shows a large, bold heading "This is a heading.". The source code is numbered from 1 to 11, showing the structure of the HTML document.

```
lesson2.html
1 <!DOCTYPE html>
2 <html>
3   <head>
4     <meta charset="utf-8">
5     <title>HTML Structure</title>
6   </head>
7   <body>
8     <h1>This is a heading.</h1>
9   </body>
10 </html>
11
```



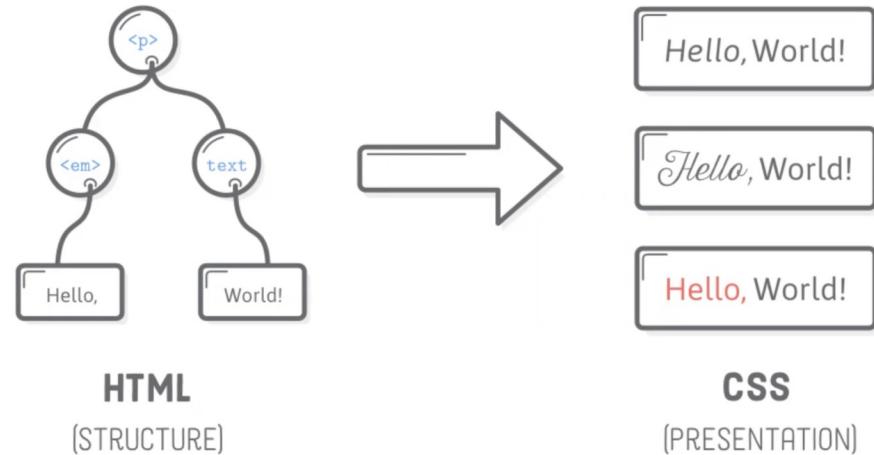
HTML in VNExpress Website



What is CSS?

- Cascading Style Sheets
- HTML is content, CSS is style

CSS defines how HTML looks like



- CSS rules have selectors and properties

HTML + CSS = ❤

Website

My biggest header

MY FIRST PARAGRAPH

MY SECOND PARAGRAPH

HTML

```
<h1>My biggest header</h1>
<p>My first paragraph</p>
<p class="red-block">My second paragraph</p>
```

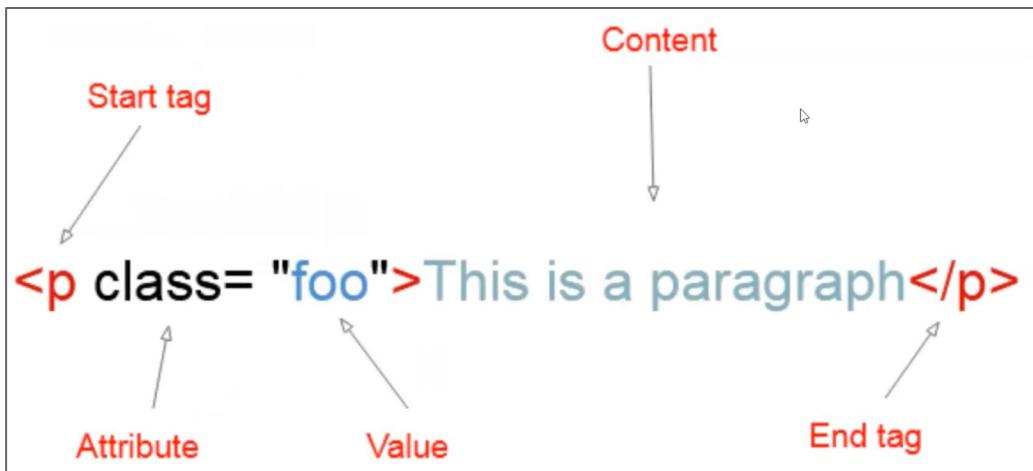
CSS

```
.red-block {
    color: red;
}
```

```
p {
    text-transform: uppercase;
}
```

Another Example of CSS

Define a CSS class “nickname” for your HTML content:



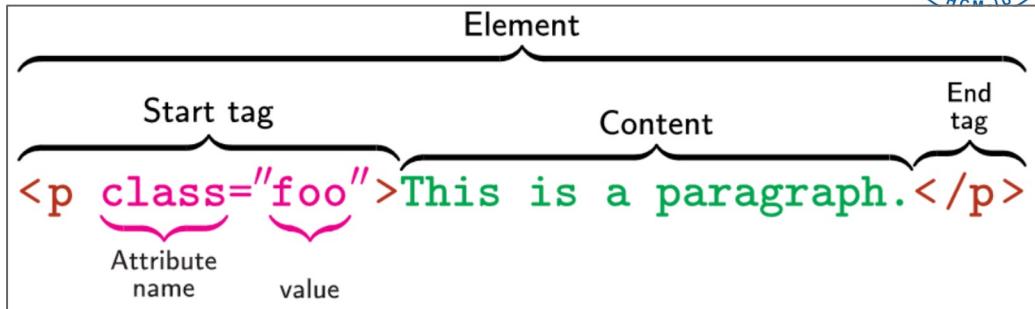
Style rules for CSS class name or HTML tag name:

```
.foo{  
    color: red;  
}  
  
p {  
    font-size: 30px;  
}
```

HTML Structure & CSS Selector

🤔 Can you guess which CSS selector targets which HTML element?

- div
- div[my_attribute]
- div[my_attribute="jerry"]
- div[new_attribute="charles"]
- p
- div[my_attribute="jerry"] div
- .my_class_name
- .my_class_name h1



Example:

```
<div my_attribute="tom">
  <div my_attribute="jerry" new_attribute="charles">
    <p>Here is a paragraph</p>
    <div class="my_class_name">
      <h1>Here is the biggest header</h1>
    </div>
  </div>
</div>
```

JSoup - Java HTML Parser



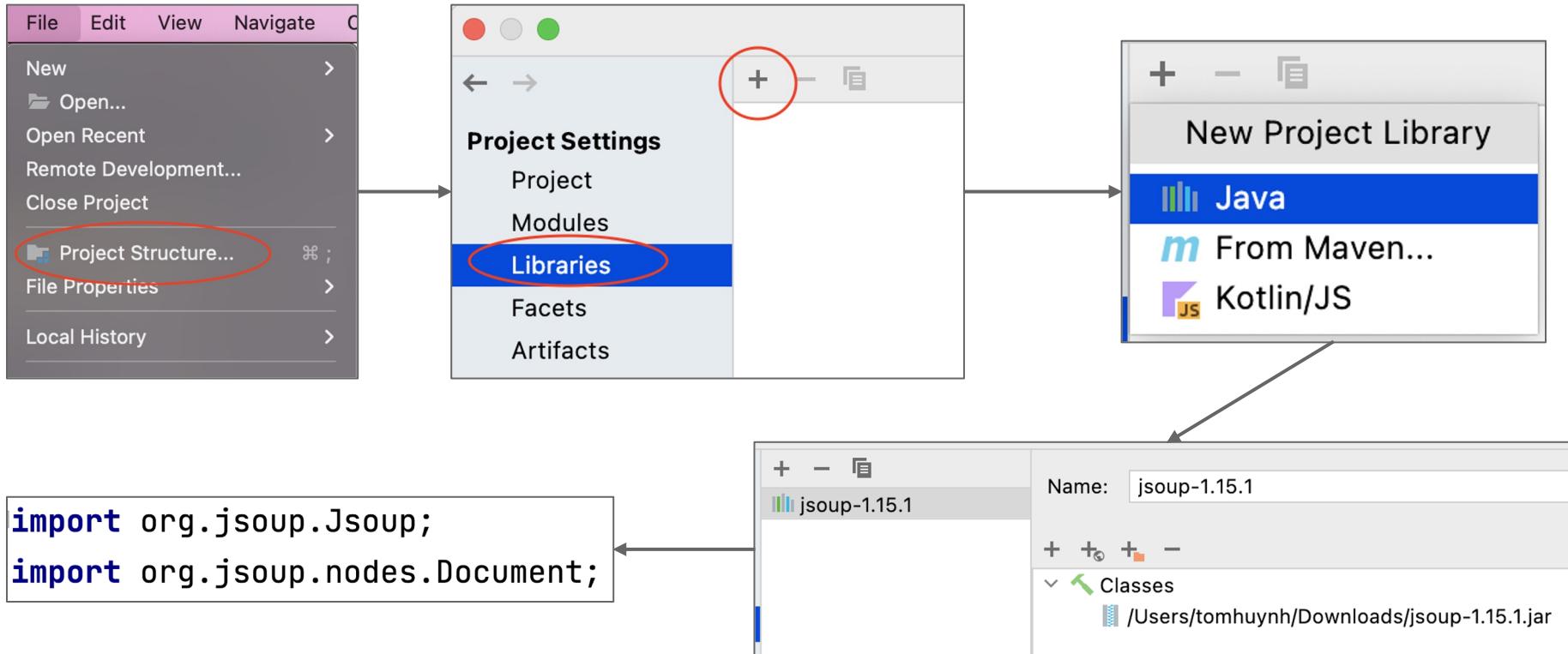
- Very convenient Java library for fetching URLs and extracting and manipulating data.
- Find and extract data using **CSS selectors**.



Setup JSoup for your project



- Download JSoup jar library on <https://jsoup.org/download>
- To use JSoup, add its jar library to your project via IDE:



JSoup Workflow



1. Get the whole HTML document response from the server via a URL:
 - a. Document doc = Jsoup.connect(url).get()

1. Select specific HTML element(s) we want to scrap from the document
 - a. Elements items = doc.select("CSS Selector Query")
 - i. You need to loop through Elements to process each Element
 - b. Element item = doc.selectFirst("CSS Selector Query")
 - i. No need to loop since you only have one Element

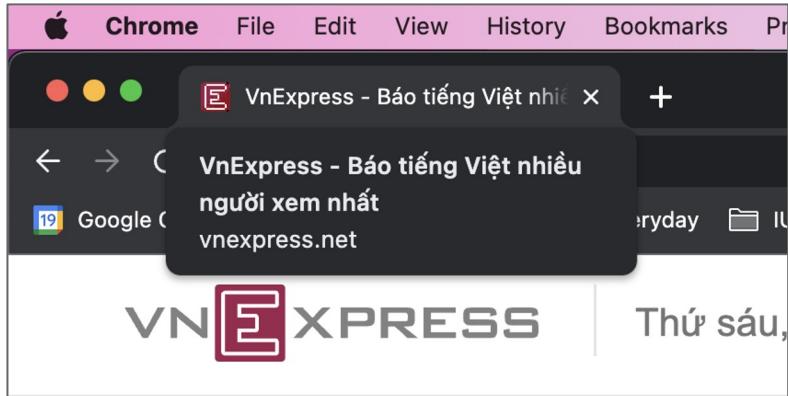
1. For each Element, extract content inside of the selected HTML element
 - a. item.text()
 - b. item.attr("attribute-name")

Super Basic Web Scraping 101



```
import org.jsoup.Jsoup;
import org.jsoup.nodes.Document;
import java.io.IOException;

public class Main {
    public static void main(String[] args) {
        Document doc = null;
        try {
            doc = Jsoup.connect(url: "https://vnexpress.net/").get();
        } catch (IOException e) {
            throw new RuntimeException(e);
        }
        String title = doc.title();
        System.out.println("title is: " + title);
    }
}
```



Output:

```
title is: VnExpress - Báo tiếng Việt nhiều người xem nhất
```

How to use it safely and legally

- A website should have a file “robots.txt” to show which links are
 - allowed to scrap
 - not allowed to scrap
- Here is an example “robot.txt”



VNExpress Scraping Exercise



VNEXPRESS | Thứ sáu, 20/5/2022 | Mới nhất | International | Tim kiếm

Thời sự Góc nhìn Thế giới Video Podcasts Kinh doanh Khoa học Giải trí Thể thao Pháp luật Giáo dục Sức khỏe Đời sống Du lịch Số hóa


Chuyển công du của Thủ tướng 'tạo đà quan hệ Việt - Mỹ'
Hơn 60 hoạt động của Thủ tướng Phạm Minh Chính trong 7 ngày công du Mỹ giúp quan hệ song phương có thêm đà tiến trong nhiều lĩnh vực, theo quan chức Mỹ và giới chuyên gia.
"Những chuyến thăm cũ, 56' trước

Ủy ban chứng khoán có người điều hành thay ông Trần Văn Dũng
Ông Trần Văn Dũng vừa bị Bộ Tài chính cách chức Chủ tịch Ủy ban chứng khoán và Thứ trưởng Tài chính Nguyễn Đức Chi được giao trực tiếp điều hành thay.

Từ chối rời sân, thủ môn Malaysia thành tội đồ
PHÚ THỌ- Không chấp nhận yêu cầu từ Ban huấn luyện, thủ thành Azri Ghani tiếp tục ở lại thi đấu và mắc sai lầm khiến Malaysia thua Việt Nam ở bán kết SEA Games 31.  67

Steve Darby: 'Tiến Linh cần được dạy lại cách sút'

Trong bài viết cho VnExpress, cựu HLV Thái Lan cho rằng Việt Nam thắng Malaysia ở bán kết SEA Games 31 nhờ tính kỹ lưỡng nhưng tiền đạo Tiến Linh cần một số đổi.  541

Báo Malaysia: 'Việt Nam ở đẳng cấp khác'

Tờ Kosmo cho rằng với sự chênh lệch rõ rệt, Malaysia thực sự đáng khen khi chỉ chịu thua Việt Nam ở hiệp phụ trận bán kết SEA Games 31 tối 19/5.  250

Kinh doanh Quốc tế Doanh nghiệp Chứng khoán Bất động sản Bảo hiểm ...

Nhà đầu tư nghiệp dư thua lỗ vì chứng khoán Mỹ
Khi Wall Street lao dốc đầu năm nay, hàng triệu nhà đầu tư cá nhân Mỹ kiếm được tiền suốt 2 năm qua phải tìm cách hâm phanh.
Gần 400 m2 biệt thự
Thị trường nhận mức phản hồi khi một dự

Nỗi sợ lạm phát kèm tăng trưởng chậm lan khắp thế giới
Chuyên gia HSBC: Cần cẩn trọng với 'vũ trụ ảo'
HAGL bồi chi giao 5 tháng

[Your Fun Game]

- Open <https://vnexpress.net/> on Chrome Browser
- Open Inspector (right click -> Inspector or F12)
- Your task: Find out there is any pattern (tag name, CSS class name) of articles in this page so that we can use these information to scrap all the title, description and article link of all articles of the homepage!

Let's scrap VNExpress!

VNEXPRESS | Thứ ba, 24/5/2022

Mới nhất | International | Tìm kiếm | Đăng nhập

Thời sự | Góc nhìn | Thế giới | Video | Podcasts | Kinh doanh | Khoa học | Giải trí | Thể thao | Pháp luật | Giáo dục | Sức khỏe | Đời sống | Du lịch | Số hóa | Xe | Ý kiến | Tâm sự | Hài | Tất cả



Biển động toàn cầu phủ bóng Diễn đàn Kinh tế Thế giới

Sau hai năm gián đoạn, Diễn đàn Kinh tế Thế giới (WEF) tổ chức trở lại tại Davos trong bối cảnh nhiều giá trị họ xây dựng dang rạn nứt.

Trong nửa thế kỷ qua, Klaus Schwab, Nhà sáng lập Diễn đàn Kinh tế Thế giới ...

2h trước

Góc nhìn

Nói với con về bất bình đẳng

Những tổn thương về sự bất bình đẳng hàn sâu vào lòng xã hội, đặc biệt sau Covid-19.

Lang Minh

TP HCM đề nghị thảo gỡ 5 vướng mắc gói hỗ trợ thuế nhà

Lo ngại bỏ lọt nhóm thu hưởng, phương thức chi trả, mức xác định thời gian thuê nhà chưa rõ ràng, Sở Lao động, Thương binh và Xã hội thành phố đề nghị thảo gỡ.

Ngày thứ 89 chiến sự Ukraine: Ông Zelensky nói xung đột đứng trước bước ngoặt

Tổng thống Zelensky kêu gọi tăng trừng phạt Nga và dấy mạnh viễn vọng Ukraine, sau khi tiết lộ tên thật quân sự lớn nhất sau ba tháng xung đột.

Quan họ Bắc Ninh thay lời tạm biệt SEA Games 31

HÀ NỘI – Các liên anh, liên chi hát quan họ, mới nghệ sĩ ruồng bạn trầu tém cánh phương, trong lễ bế mạc SEA Games 31, tối 23/5. [\[xem\]](#)

Hơn nửa đội hình vô địch SEA Games 31 dự giải U23 châu Á

HÀ NỘI – Tân HLV Gong Oh-kyun chia sẻ về sự kiện HLV và SEA Games và danh

Kinh doanh

Quốc tế | Doanh nghiệp | Chứng khoán | Bất động sản | Bảo hiểm | E-Commerce 4.0



Dòng vốn ngoại đang rời Trung Quốc ra sao?

Vốn ngoại rời Trung Quốc năm nay có thể gấp đôi 2021 khi nhà đầu tư e ngại các chính sách kinh tế, chính sách, đối ngoại của Bắc Kinh.

Ngân sách năm 2020 thảm hụt hơn 216.400 tỷ đồng

Mức赤chết ngân sách sau quyết toán năm 2020 là hơn 216.405 tỷ đồng (gần 9.3 tỷ USD), thấp hơn nhiều ...

- Biden: Mỹ có thể tránh được suy thoái
- Những lợi thế của bất động sản quân 7
- Giới bất động sản đón sóng đầu tư ở Tân Uyên



VNExpress Homepage Scraping Result

Title: Đề xuất gần 113 dự án dùng vốn từ gói phục hồi kinh tế

Article: <https://vnexpress.net/de-xuat-gan-113-du-an-dung-von-tu-goi-phuc-hoi-kinh-te-4466119.html>

Description: Số dự án, nhiệm vụ dùng vốn từ gói phục hồi kinh tế được đề xuất còn 113, giảm khoảng một phần ba so với con số đưa ra cách đây hơn một tháng.

Title: Độ đèn tuổi 18

Article: <https://vnexpress.net/do-den-tuoi-18-4465661.html>

Description: Tôi không phản đối casino nhưng bây giờ chưa phải lúc để người Việt 18 tuổi vào chơi.

Title: Sóng ngầm cắt lỗ căn hộ của nhà đầu tư F0

Article: <https://vnexpress.net/song-ngam-cat-lo-can-ho-cua-nha-dau-tu-f0-4465898.html>

Description: TP HCM vài tháng nay, anh Đinh rao bán lỗ 100 triệu đồng căn hộ một phòng ngủ, giá 1,7 tỷ đồng đang xây móng, song chẳng ai hỏi mua.

Title: Nga tuyên bố sắp kiểm soát toàn bộ tinh miền đông Ukraine

Article: <https://vnexpress.net/nga-tuyen-bo-sap-kiem-soat-toan-bo-tinh-mien-dong-ukraine-4466155.html>

Description: Bộ trưởng Shoigu thông báo lực lượng Nga và phe ly khai sắp kiểm soát toàn bộ tinh Lugansk ở miền đông Ukraine sau gần ba tháng chiến sự.

Missing Article Title!

Article: /the-thao/sea-games-31?utm_source=VnExpress&utm_medium=Box&utm_campaign=SeaGames31_PC&utm_content=SEAGames31

Missing Article Description!

Title: Nghệ sĩ Phượng Loan gặp tai nạn giao thông

Article: <https://vnexpress.net/nghe-si-phuong-loan-gap-tai-nan-giao-thong-4466095.html>

Description: Nghệ sĩ cải lương Phượng Loan ngã từ trên xe máy khi đi hát đám cưới, được chẩn đoán xuất huyết não, gãy xương má...

Put The Result Into a Nice Table

	title	url	image_url	description
0	Đại sứ Mỹ: Chiến lược với châu Á sẽ không đổi ...	https://vnexpress.net/dai-su-my-chien-luoc-voi...	https://vcdn1-vnexpress.vnecdn.net/2020/09/04/...	Đại sứ Mỹ tại Việt Nam Daniel Kritenbrink cho ...
1	Phương án sửa Bảo tàng Mỹ thuật TP HCM	https://vnexpress.net/phuong-an-sua-bao-tang-m...	https://vcdn1-vnexpress.vnecdn.net/2020/09/24/...	Chủ đầu tư cao ốc từ giác Bến Thành gửi Sở Văn...
2	Việt Nam mong muốn phát huy quan hệ hợp tác to...	https://vnexpress.net/viet-nam-mong-muon-phat-...	https://vcdn1-vnexpress.vnecdn.net/2020/09/25/...	Tổng Bí thư, Chủ tịch nước Nguyễn Phú Trọng ch...
3	Bỗng dung mắc nợ ngân hàng dù không vay vốn	https://vnexpress.net/bong-dung-mac-no-ngan-ha...		Nhiều khách hàng bỗng dung phát hiện mắc nợ c...
4	Cô dâu Việt 'đổi đời' nhờ mẹ chồng Tây	https://vnexpress.net/co-dau-viet-doi-doi-nho...	https://vcdn1-giadinh.vnecdn.net/2020/09/25/11...	AnhĐang tuyệt vọng vì bạn trai cắt đứt mọi liê...
5	Kịch bản Trump từ chối rời Nhà Trắng gây lo ngại	https://vnexpress.net/kich-ban-trump-tu-choi-r...	https://vcdn1-vnexpress.vnecdn.net/2020/09/24/...	Trump mới đây cho biết ông không hứa rời Nhà T...
6	Giá xăng ngày mai có thể giảm tiếp	https://vnexpress.net/gia-xang-ngay-mai-co-the...	https://vcdn1-kinhdoanh.vnecdn.net/2020/09/25/...	Thị trường thế giới đi xuống so với chu kỳ tru...
7	Cháy khách sạn, một người chết	https://vnexpress.net/chay-khach-san-mot nguoi...	https://vcdn1-vnexpress.vnecdn.net/2020/09/25/...	TP HCMTiến vào khu vực lửa đang bùng lên ở tần...
8	Chuẩn bị khởi công 3 dự án cao tốc Bắc Nam	https://vnexpress.net/chuan-bi-khoi-cong-3-du...	https://vcdn1-vnexpress.vnecdn.net/2020/09/25/...	Các dự án Mai Sơn - quốc lộ 45, Vĩnh Hảo - Pha...
9	Nga - Trung công kích Mỹ tại Liên Hợp Quốc	https://vnexpress.net/nga-trung-cong-kich-my-t...	https://vcdn1-vnexpress.vnecdn.net/2020/09/25/...	Đại sứ Trung Quốc tại Liên Hợp Quốc cáo buộc M...
10	Bayern đoạt Siêu Cup châu Âu	https://vnexpress.net/bayern-doat-sieu-cup-cha...	https://vcdn1-thethao.vnecdn.net/2020/09/25/ba...	HungaryThắng ngược Sevilla 2-1 ở hiệp phụ tối ...
11	Nhà phân phối mới xe Nissan tại Việt Nam là ai?	https://vnexpress.net/nha-phan-phoi-moi-xe-nis...	https://vcdn1-vnexpress.vnecdn.net/2020/09/25/...	Công ty TNHH Phát triển Công nghiệp Ôtô Việt N...
12	Nạn nhân bị cây đè tử vong sau mổ	https://vnexpress.net/nan-nhan-bi-cay-de-tu-vo...	https://vcdn1-suckhoe.vnecdn.net/2020/09/25/ca...	TP HCMNạn nhân 35 tuổi, bị cây đổ trong căn mư...
13	300 năm người Thái áp út tham vọng 'kênh đào Pa...	https://vnexpress.net/300-nam-nguo-thai-ap-u...		Cứ vài năm, ý tưởng đào con kênh qua Eo đất Kr...
14	Loạt dịch vụ Google gấp sự cố	https://vnexpress.net/loat-dich-vu-google-gap...	https://vcdn1-sohoa.vnecdn.net/2020/09/25/goog...	Gmail, Drive, YouTube và nhiều dịch vụ Google ...
15	22 ngày không lây nhiễm nCoV cộng đồng	https://vnexpress.net/22-ngay-khong-lay-nhiem-...		Bộ Y tế sáng 25/9 không ghi nhận thêm ca dương...

Your Homework

- Scrap all the links in the footer of English website of Vnexpress (<https://e.vnexpress.net/>):

News	Business	Travel	Life	Sports	Video	Perspectives	VLight	Contact us
	Economy	Places	Trend				Buy VnExpress Photos	Get Newsletter
	Industries	Food	Culture					Advertise with us:
	Companies		Style					 090 293 9644
	DataSpeaks							 Click to follow

- Return a dictionary which the key is the name of the menu item and the value is the link/url of that menu item. Thus, there should be 22 items in the dictionary looking like this:

```
{ "News" : "/news/news",
  "Business" : "/news/business",
  "Economy" : "/news/business/economy",
  ...
  "Contact us" : "/contactus",
  "Get Newsletter" : "/newsletter",
```

Italian Recipe Scrapping Homework

<https://foodnetwork.co.uk/italian-family-dinners/>

The screenshot shows a website header with navigation links: Recipes, Veganuary, Inspiration, TV Guide, Stream on discovery+, Shows, Chefs, Competitions. Below this is a large image of a plate of gnocchi with sage leaves. To the left, the text "88 Italian Family Dinners" and "We've hand picked these delicious Italian meals that were made for sharing." are displayed. At the bottom, five recipe cards are shown:

- Linguine alla Parmigiana** (Easy) - Luscious linguine tossed with wafer-thin slices of Parma ham and tender peas.
- Parmigiano Reggiano and Truffle Tagliolini** (Easy) - For the fresh pasta, mix all the ingredients, cover the dough with cling film and leave to rest in the fridge, about 30 minutes. 2. After 1 hour minimum unwrap the dough...
- Parmigiano Reggiano, Pumpkin and Porcini Risotto** (Easy) - Use dried porcini mushrooms to add an intense depth of flavour to this classic.
- Gorgonzola Piccante Pizza with Courgettes and Balsamic Onions** (Easy) - Preheat the oven to 200°C. Melt the butter in a heavy based saucepan and add the onions, sugar and vinegar. Cook over a medium heat for 15-20 minutes...
- Gorgonzola Salad with Pistachios and Balsamic Vinaigrette** (Easy) - This light, fresh salad is a must-have side dish.

Scrap all Italian recipes, which should have:

- Recipe title
- Short description
- Preparation time
- Cook time
- Number of serves



A Fun Project with Manga

MANGA DOOM

Daily updates, fastest & latest

Search Manga

HOME MANGA DIRECTORY POPULAR MANGA LATEST CHAPTERS ADVANCED SEARCH ALL GENRES

POPULAR UPDATES

Martial Peak - 2252
 Tower of God - 543
 The Gamer - 424
 The Legend Of The dem...
 Rebirth Of the Urban I...
 Other World Warrior - ...

• • • • •

LATEST MANGA UPDATES

Soul Land IV - The Ultimate Combat

Soul Land IV - The Ultimate Combat - 314
Soul Land IV - The Ultimate Combat - 313.5

9 Hours, 11 Minutes

The Lord of No Boundary

The Lord of No Boundary - 260
The Lord of No Boundary - 259
The Lord of No Boundary - 258
The Lord of No Boundary - 257

10 Hours, 27 Minutes

<http://mngdoom.chatango.com>

Wed Mar 30, 8:43:58pm
sage: AGREED! This site is just to make sure im
not missing anything.

athx: The Beginning After the End...that
ending...wow

moderator: Yes

Sat Apr 2, 3:15:13am

ss: demon slayer op

Sun Apr 3, 9:45:44pm

anon3485: hello

Wed Apr 13, 8:57:02pm

sussyamogus: no

Fri Apr 15, 8:17:49pm

anon4420: Po

Sat Apr 16, 7:09:54am

Our Challenge



- Let's scrap one of the best manga websites, "MangaDoom"!

<https://www.mngdoom.com/>

MANGA DOOM
Daily updates, fastest & latest

HOME MANGA DIRECTORY POPULAR MANGA LATEST CHAPTERS ADVANCED SEARCH ALL GENRES

POPULAR MANGA

1. BATTLE THROUGH THE HEAVENS

Alternative Name: Dou Po Cang Qiong, Fights Break Sphere
Status: -
Categories: Action, Adventure, Comedy, Romance, Shounen, Supernatural
Type: Chinese
Author: Tian Can Tu Dou
Artist: N/A
Total Views: 30,345,011

2. TALES OF DEMONS AND GODS

Alternative Name: 鬼神記
Status: Ongoing
Categories: Action, Adventure, Fantasy, Martial Arts, Romance, Shounen
Type: Chinese
Author: Mad Snail
Artist: Mad Snail
Total Views: 29,706,217

3. THE GOD OF HIGH SCHOOL

Alternative Name: 강오아리스, God of Highschool, GoH, The God of Highschool
Status: -
Categories: Action, Adventure, Comedy, Martial Arts, Shounen
Type: Korean
Author: PARK Yong Je
Artist: PARK Yong Je
Total Views: 29,038,768

4. KINGDOM

Alternative Name: キングダム: 王者天下
Status: Ongoing
Categories: Action, Historical

Search Manga

LOGIN SIGN UP

Your email address
Password
[Lost password?](#)

POPULAR MANGA

1. BATTLE THROUGH THE HEAVENS
Battle Through the Heavens - 640
21/10/2021 30,345,007 views

2. TALES OF DEMONS AND GODS
Tales of Demons and Gods - 396.5
12/03/2022 29,706,207 views

3. THE GOD OF HIGH SCHOOL
The God of High School - 541
30/04/2022 29,038,765 views

4. KINGDOM
Kingdom - 719
13/05/2022 24,180,590 views

Chapter list

MANGA DOOM
Daily updates, fastest & latest

HOME MANGA DIRECTORY POPULAR MANGA LATEST CHAPTERS ADVANCED SEARCH ALL GENRES

ONE PIECE

Alternative Name: ワンピース, 海賊王, 원피스
Status: -
Categories: Action, Adventure, Comedy, Drama, Fantasy, Martial Arts, Mystery, Shounen
Type: Japanese
Author: Oda Eiichiro
Artist: Oda Eiichiro
Total Views: 21,250,523

Share

Greatness, Glory, Gold. A Pirate Named Gold Roger also known as the Pirate King has conquered this all. He was executed with an unknown reason but before he died he had revealed his last word about The legendary Treasure named One Piece which was hidden in the Grand Line. 22 years after his Death, A Pirate named Monkey D. Luffy appeared and has only one Goal, To Become the Next "Pirate King" and Find The Treasure "One Piece". This Start The Never Ending Adventure. Monkey D. Luffy, inspired by his childhood hero "Red-Haired" Shanks, sets out on a journey to find the legendary treasure One Piece, to become the new Pirate King, and after eating the devil's fruit gains the power to do so. To accomplish this, he must reach the end of the most deadly and dangerous ocean: The Grand Line. But first he must find a crew.

LOGIN SIGN UP

Your email address
Password
[Lost password?](#)

POPULAR MANGA

1. BATTLE THROUGH THE HEAVENS
Battle Through the Heavens - 640
21/10/2021 30,345,007 views

2. TALES OF DEMONS AND GODS
Tales of Demons and Gods - 396.5
12/03/2022 29,706,207 views

3. THE GOD OF HIGH SCHOOL
The God of High School - 541
30/04/2022 29,038,765 views

4. KINGDOM
Kingdom - 719
13/05/2022 24,180,590 views

Chapter list

Let's pick a best manga to scrap them!



- “One Piece” is my favourite manga of all time! Let’s scrap the whole manga!



Let's do it together!



Scrap our favorite manga!

MANGA DOOM

Daily updates, fastest & latest

HOME

MANGA DIRECTORY

POPULAR MANGA

LATEST CHAPTERS

ADVANCED SEARCH

ALL GENRES

ONE PIECE



Alternative Name: ワンピース, 海賊王, 원피스

Status:

Categories:

Action, Adventure, Comedy, Drama, Fantasy, Martial Arts, Mystery, Shounen

Type: Japanese

Author:

Oda Eiichiro

Artist:

Oda Eiichiro

Total Views: 21,260,654

Share

Greatness, Glory, Gold. A Pirate Named Gold Roger also known as the Pirate King has conquered this all. He was executed with an unknown reason but before he died he had revealed his last word about The legendary Treasure named One Piece which was hidden in the Grand Line. 22 years after his Death, A Pirate named Monkey D. Luffy appeared and has only one Goal, To Become the Next "Pirate King" and Find The Treasure "One Piece". This Start The Never Ending Adventure. Monkey D. Luffy, inspired by his childhood hero "Red-Haired" Shanks, sets out on a journey to find the legendary treasure One Piece, to become the new Pirate King, and after eating the devil's fruit gains the power to do so. To accomplish this, he must reach the end of the most deadly and dangerous ocean: The Grand Line. But first he must find a crew.

Search Manga



http://mngdoom.chatango.com

Wed Jan 26, 7:21:32am
St0n3rb0i: is there dragon ball on here

Wed Jan 26, 8:55:21am
athx: @St0n3rb0i super? yes it's on here

Thu Jan 27, 3:29:05am
panda: What do you do when u encounter Error 502

Thu Jan 27, 5:47:15am
athx: i got this error on mobile but never on my pc. maybe clear cache or try a different browser?

Thu Jan 27, 11:27:26pm
anon3223: half the time this is trash, images not showing

Sat Jan 29, 3:25:59pm
anon9951: hi i got a question

Sat Jan 29, 3:26:36pm
anon9951: where can i find the chapter of demon slayer that's just release this week?

Type here to send a message

Chatango 14 Set name

LOGIN

SIGN UP

Your email address

Password



I'm a collector myself! Web Scraping is awesome!

Volumes/Everyday.../One Piece Digital Colored/ - XnView MP

Browser Quick search

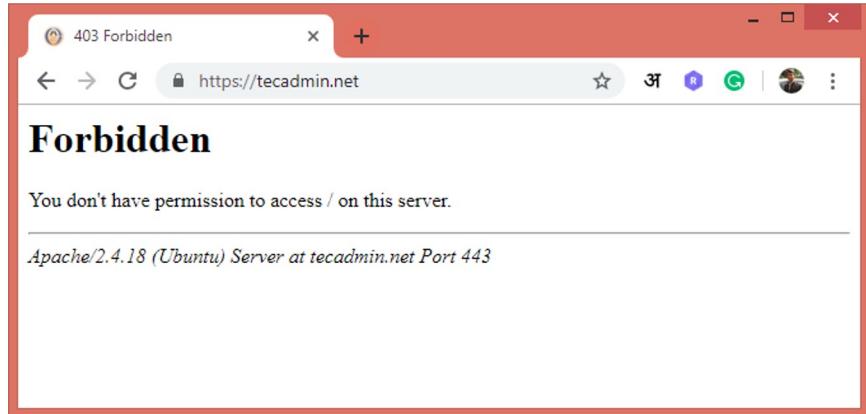
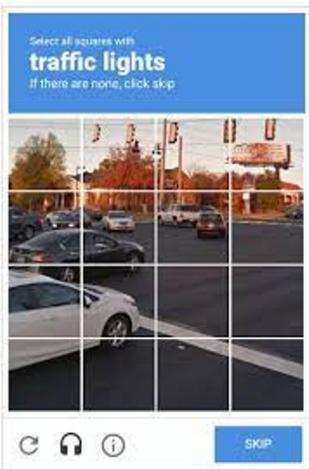
Folders Favorites Categories Filter

The screenshot shows a file browser window titled "Volumes/Everyday.../One Piece Digital Colored/ - XnView MP". The main pane displays a grid of numerous small thumbnail images, each representing a chapter from the manga "One Piece Digital Colored". The thumbnails are arranged in approximately 20 rows and 20 columns. The first few rows contain chapters numbered Chapter 001 through Chapter 050. Subsequent rows continue this pattern up to Chapter 0950. The thumbnails are blue with white icons, and the overall interface is a standard Mac OS X style.

Potential Problems & Solutions

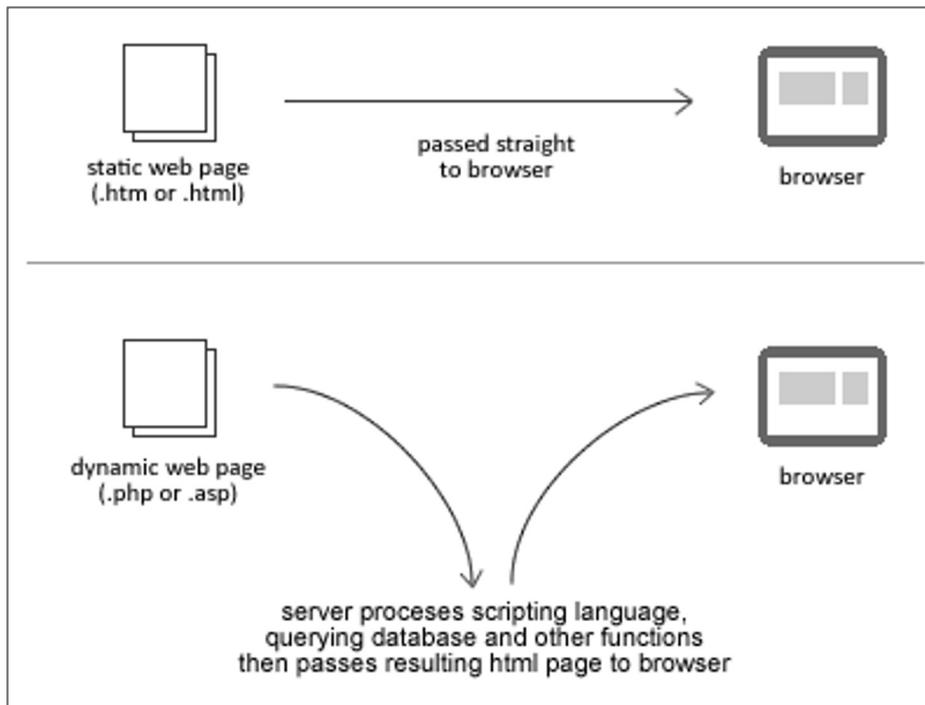
Potential Issues for Scraping

- Completely Automated Public Turing Tests (CAPTCHAs)
- IP blocking
- Geo-blocking
- Dynamic Websites



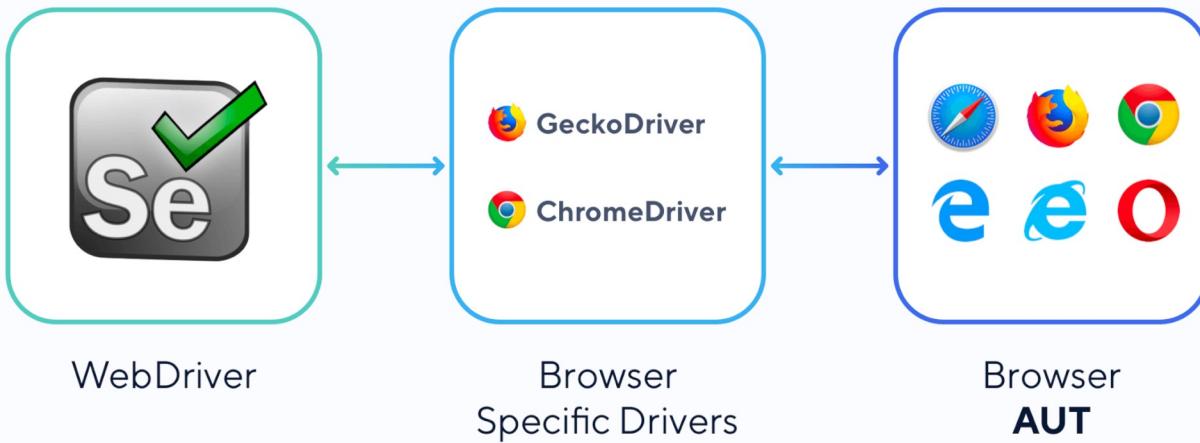
Static website vs Dynamic website

- So far, we have only learned enough to scrap static websites but no dynamic websites.



Selenium to scrape dynamic website

Selenium WebDriver Architecture



Fun Quiz

- Open popular ecommerce websites in Vietnam:
 - <https://tiki.vn/>
 - <https://www.lazada.vn/>
 - <https://shopee.vn/>
 - <https://www.sendo.vn/>
- Open your favourite website ever:
 - <https://it.hcmiu.edu.vn/>
- Open other websites:
 - <https://baomoi.com/>
 - <https://mangasee123.com/>
 - https://vi.wikipedia.org/wiki/Tr%C6%B0%E1%BB%9Dng_%C4%90%E1%BA%A1i_h%E1%BB%8Dc_Qu%E1%BB%91c_t%E1%BA%BF,_%C4%90%E1%BA%A1i_h%E1%BB%8Dc_Qu%E1%BB%91c_gia_Th%C3%A0nh_ph%E1%BB%91_H%E1%BB%93_Ch%C3%AD Minh

[Question] Which one is static and which one is dynamic? Which one can be scrapped with JSoup?

Ultimate Challenge

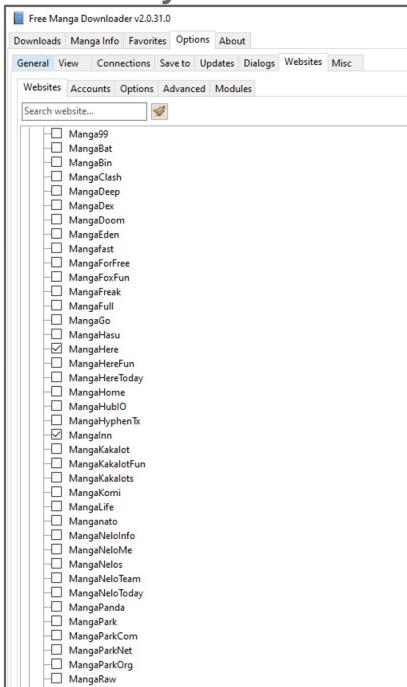
- MangaSee is the only website which got the **FULL-COLORED** chapters of big mangas like One Piece and Bleach!!!
- [Task] Can you figure out how to scrap either full-colored One Piece or Bleach?
 - <https://mangasee123.com/manga/One-Piece-Digital-Colored-Comics>
 - <https://mangasee123.com/manga/Bleach-Color>



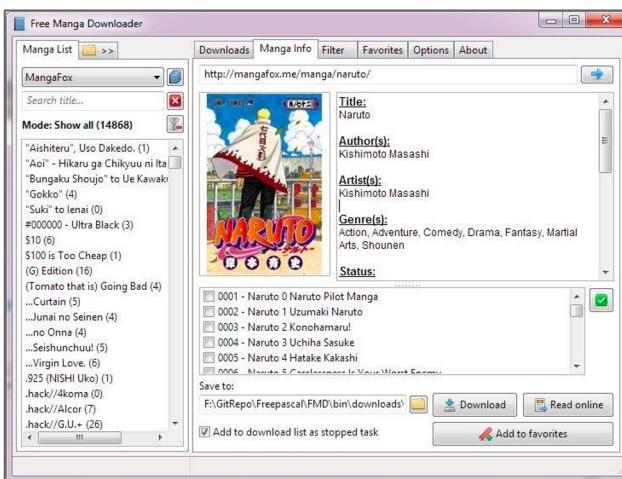
One of The Best Manga Downloader Project



- <https://github.com/riderkick/FMD>
- The Free Manga Downloader is a free open source application written in Object Pascal for managing and downloading manga from various websites.
- Based on what you have learned so far about web scraping, now you can actually start to contribute to this open source project to help the Manga community!



The screenshot shows the GitHub repository page for 'riderkick / FMD' (Public). At the top, it shows the repository name, 'Code' tab, issues (509), pull requests (6), actions, projects, wiki, security, and insights. The repository has 4 branches (master, 3rd, baseunits, config) and 214 tags. A recent commit by 'kmvi DigitalTeam' is listed: 'add external sources' (commit 84f9d02, Jan 18, 2019, 3,646 commits). Below the commit list, there are several pull requests (PRs) with their titles, descriptions, and timestamps. The PRs include: 'remove BESEN and add duktape' (4 years ago), 'MangaHere, rewrite' (3 years ago), 'delete mangalist.ini and move the rest to base.ini #1020' (4 years ago), 'mangafox, added watermark template' (5 years ago), 'DigitalTeam, add external sources' (3 years ago), 'fix NPE' (3 years ago), 'Update updater.pt_BR.po' (4 years ago), 'updater, increase retry count' (4 years ago), 'update .gitattributes' (7 years ago), and 'added slim updater' (4 years ago).



Recap



- **What is Web Scraping?**
- **Why is it important?**
- **How to use it safely and legally?**
- HTML, CSS
- Introduction to **JSoup Library**
- JSoup workflow
- VNExpress Scrapping Demo
- Manga Scrapping Demo
- Potential issues with Web Scraping

Thank you for your listening!

“Live as if you were to die tomorrow.
Learn as if you were to live forever!”

Mahatma Gandhi

