Vietnam National University of HCMC
International University
School of Computer Science and Engineering

Web Scraping

(IT069IU)

Le Duy Tan, Ph.D.

ldtan@hcmiu.edu.vn

leduytanit.com

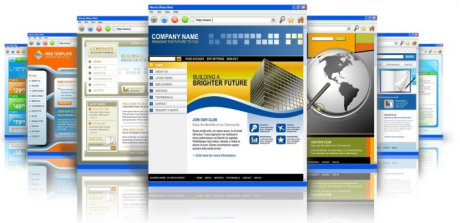
Previously,

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

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

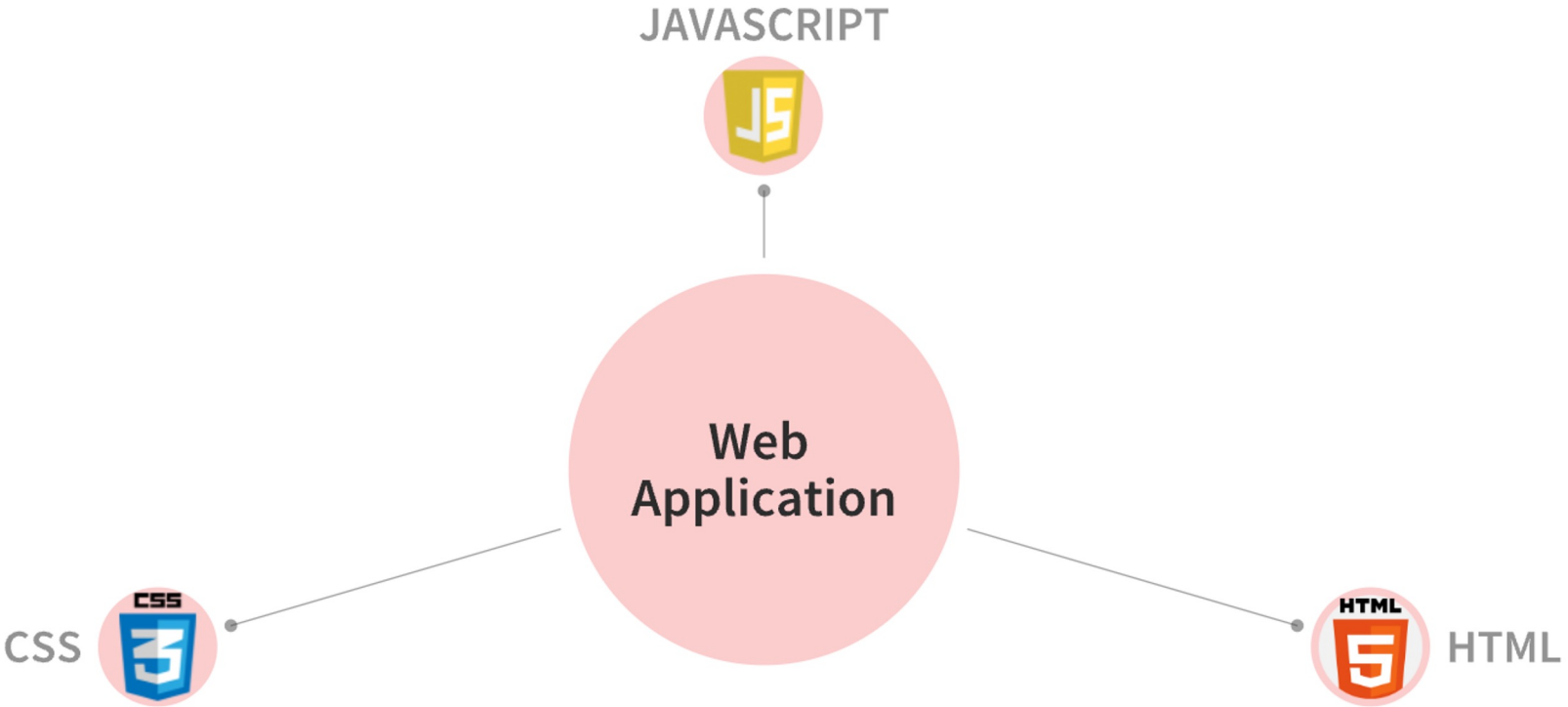
Why is Web Scraping Used?

- 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

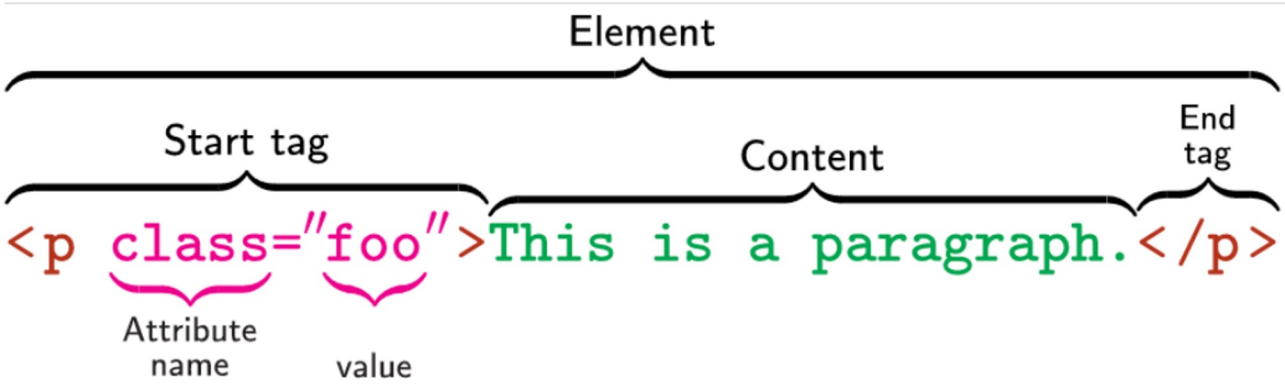
CSS

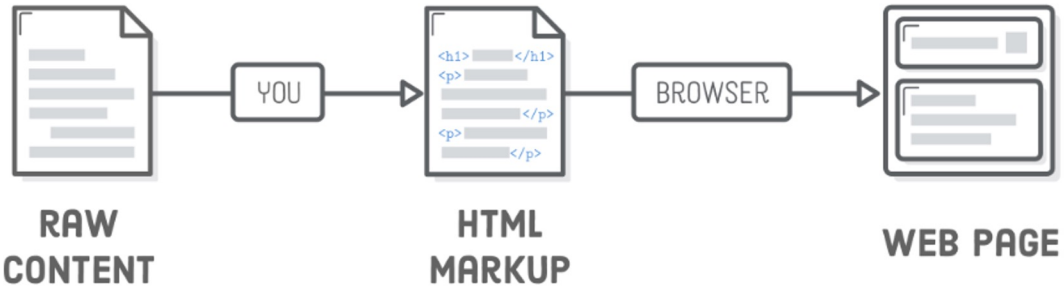
Javascript

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

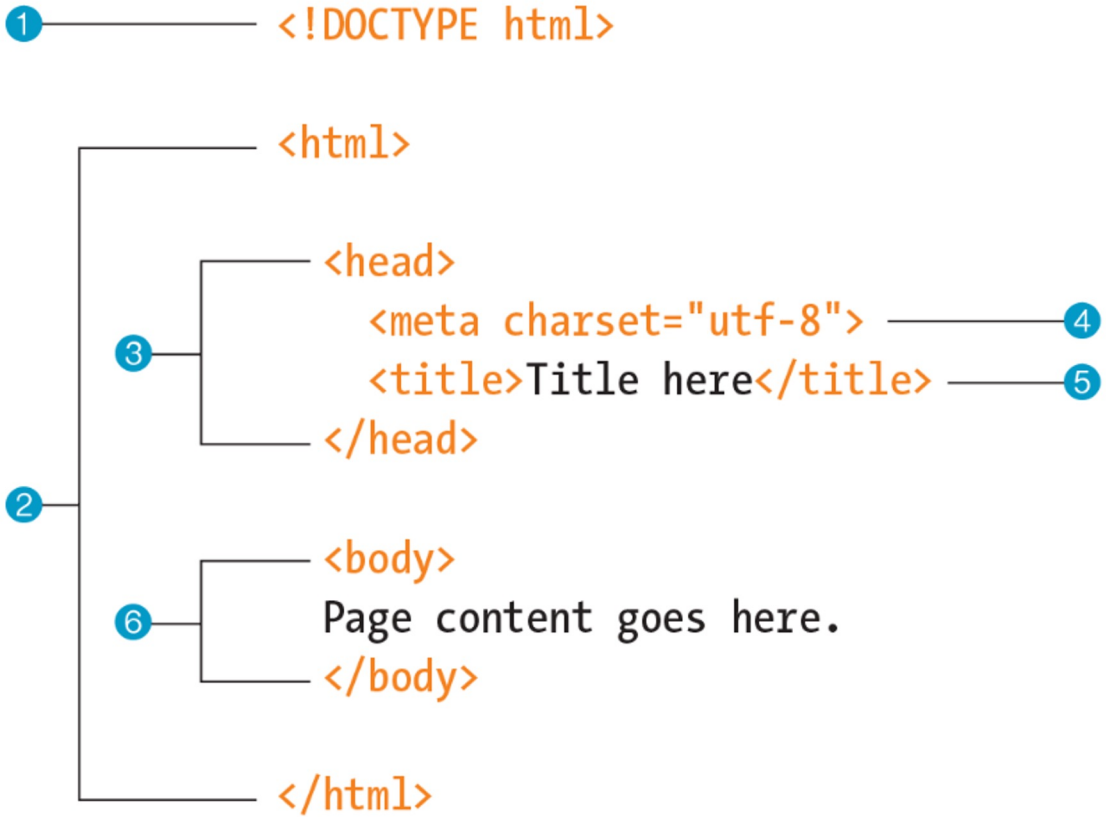




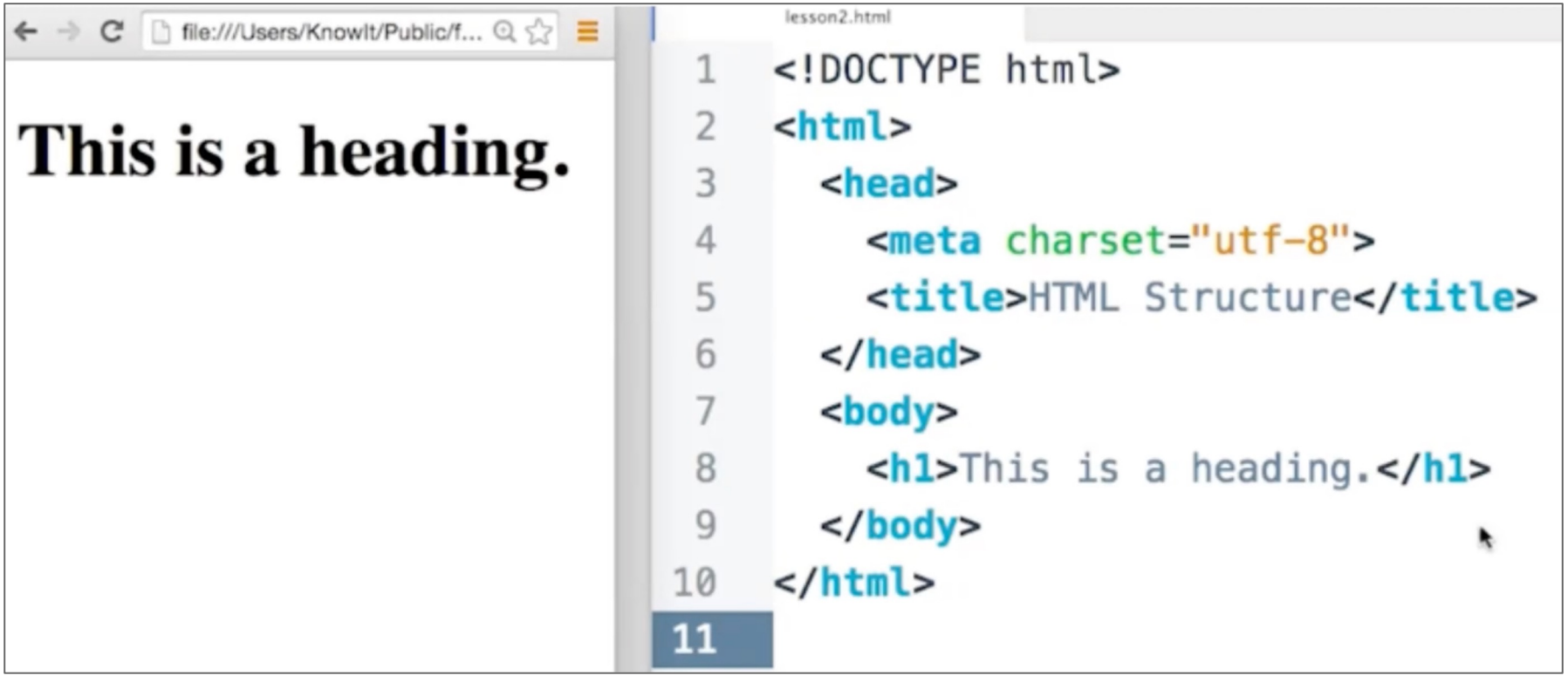
Common HTML Tags

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

Basic HTML Structure



Sample Basic HTML

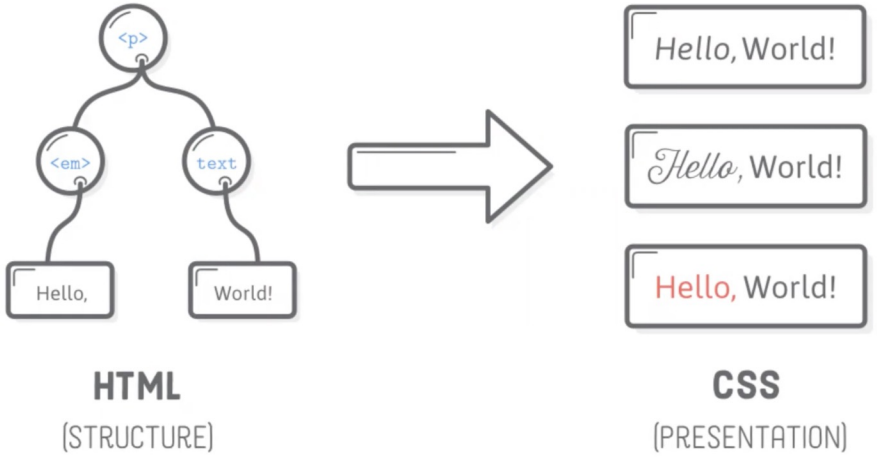


HTML in VNExpress Website

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

HTML

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

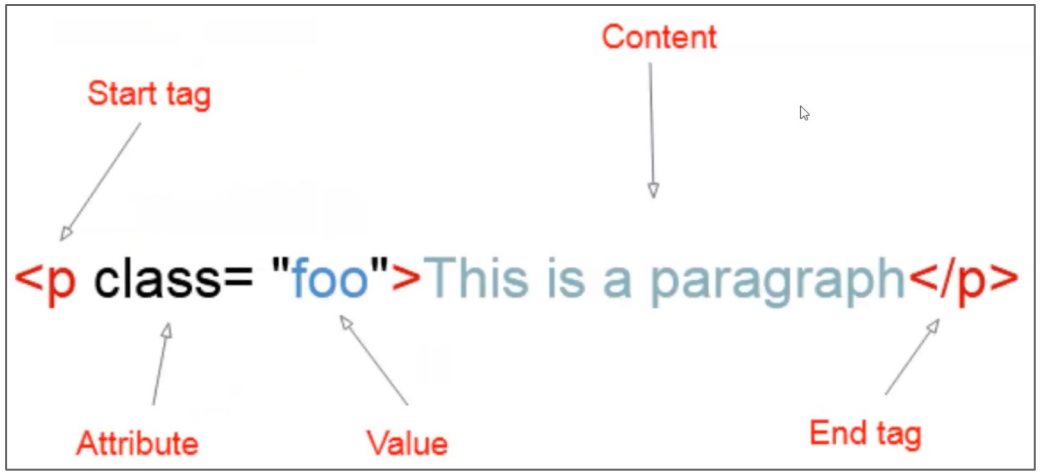
CSS

```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”]

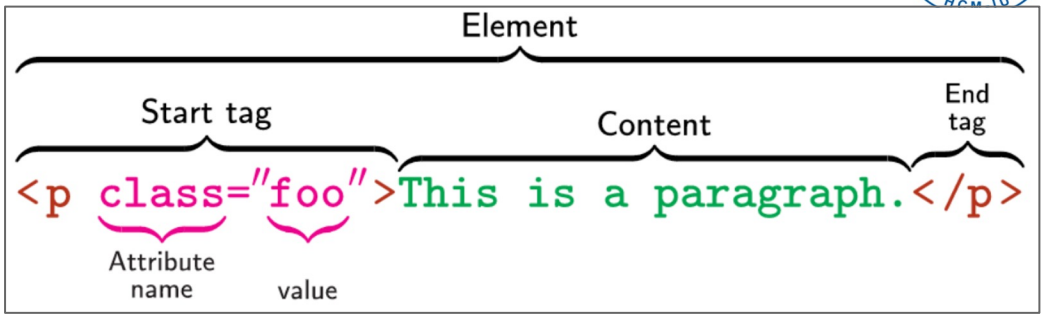
- div[new_attribute=”charles”]

- 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>
```

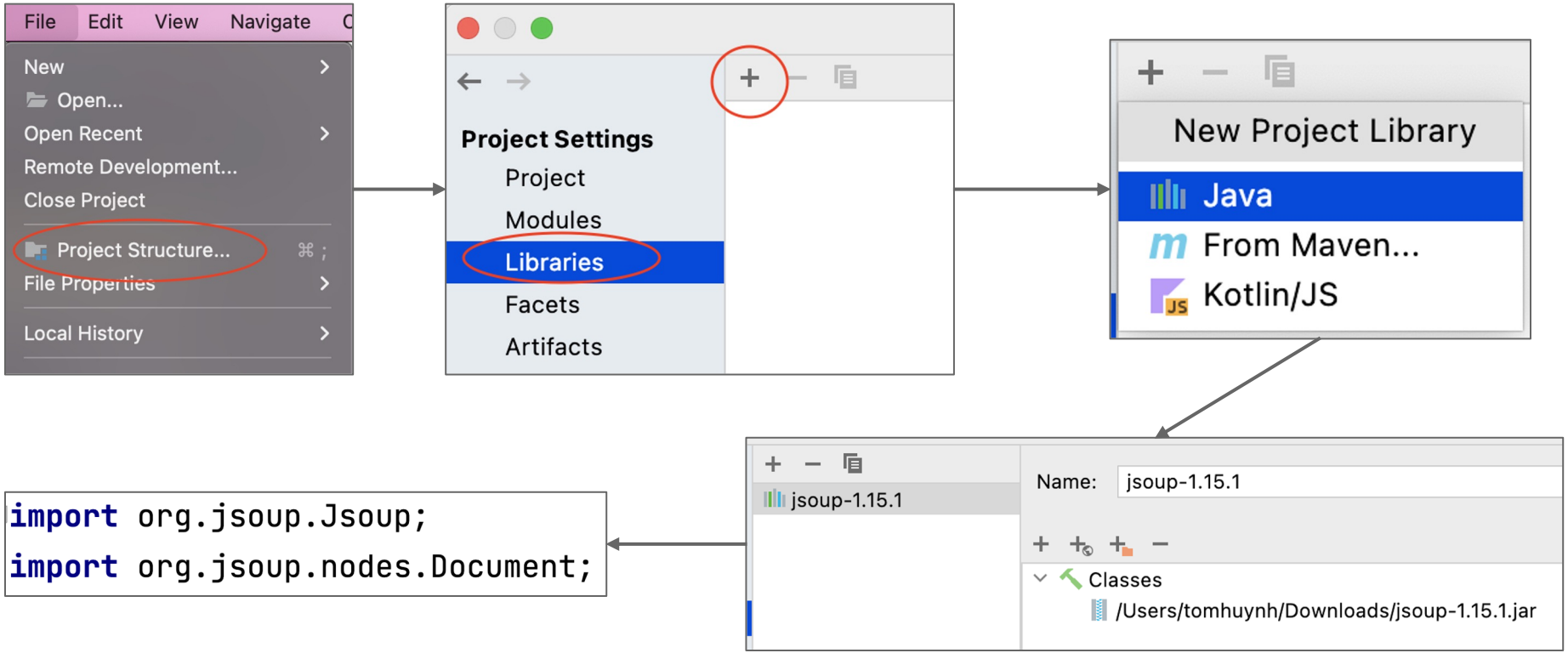
JSoup - Java HTML Parser

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

https://jsoup.org/

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

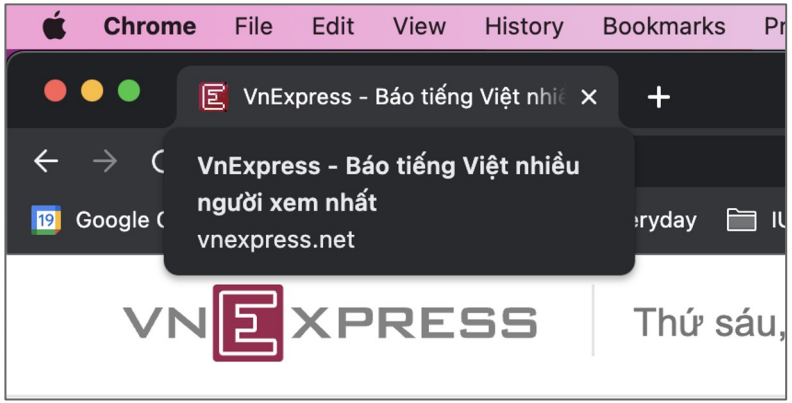
1. For each Element, extract content inside of the selected HTML element
    a. item.text()
    b. item.attr("attribute-name")

https://jsoup.org/apidocs/

Super Basic Web Scraping 101

```java
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);
    }
}
```



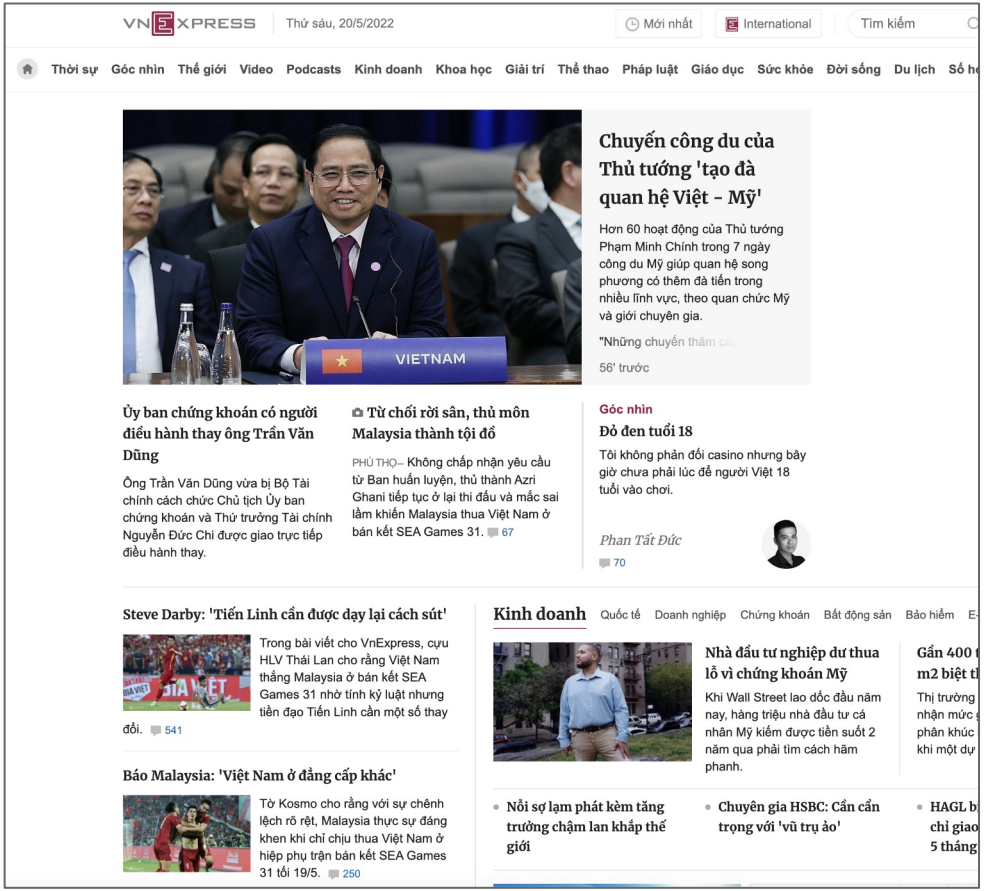
Output:

```
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”

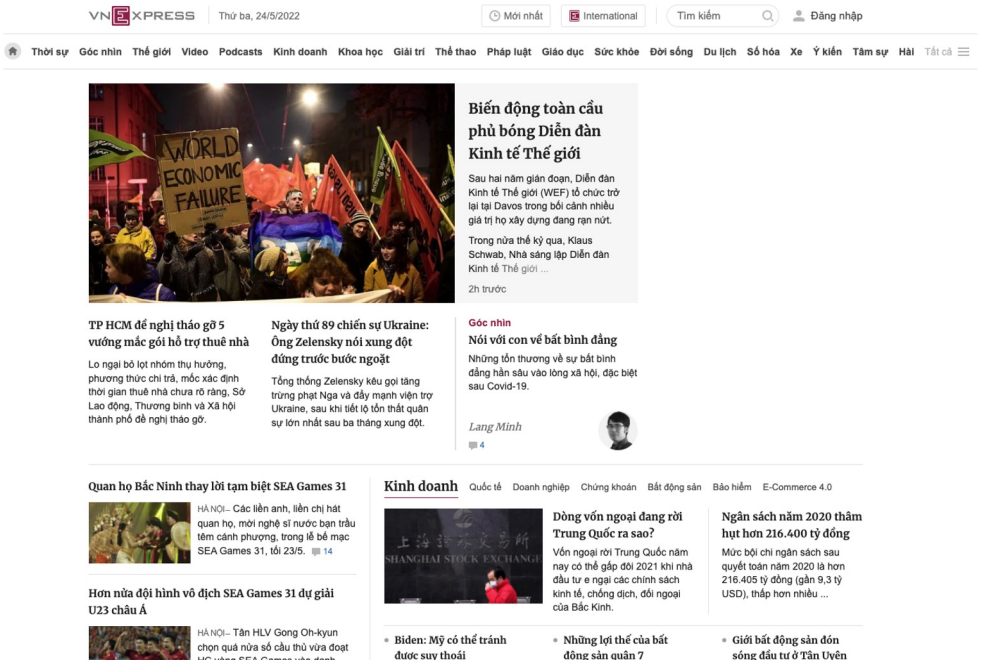
VNExpress Scraping Exercise



[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 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: Đỏ đen 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 HCMVài tháng nay, anh Điền 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 tỉnh 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 tỉnh 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 tứ 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 dưng mắc nợ ngân hàng dù không vay vốn	https://vnexpress.net/bong-dung-mac-no-ngan-ha...		Nhiều khách hàng bỗng dưng phát hiện mắc nợ cô...
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 trư...
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 cơn mư...
13	300 năm người Thái ấp ủ tham vọng 'kênh đào Pa...	https://vnexpress.net/300-nam-nguoi-thai-ap-u-...		Cứ vài năm, ý tưởng đào con kênh qua Eo đất Kr...
14	Loạt dịch vụ Google gặp 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

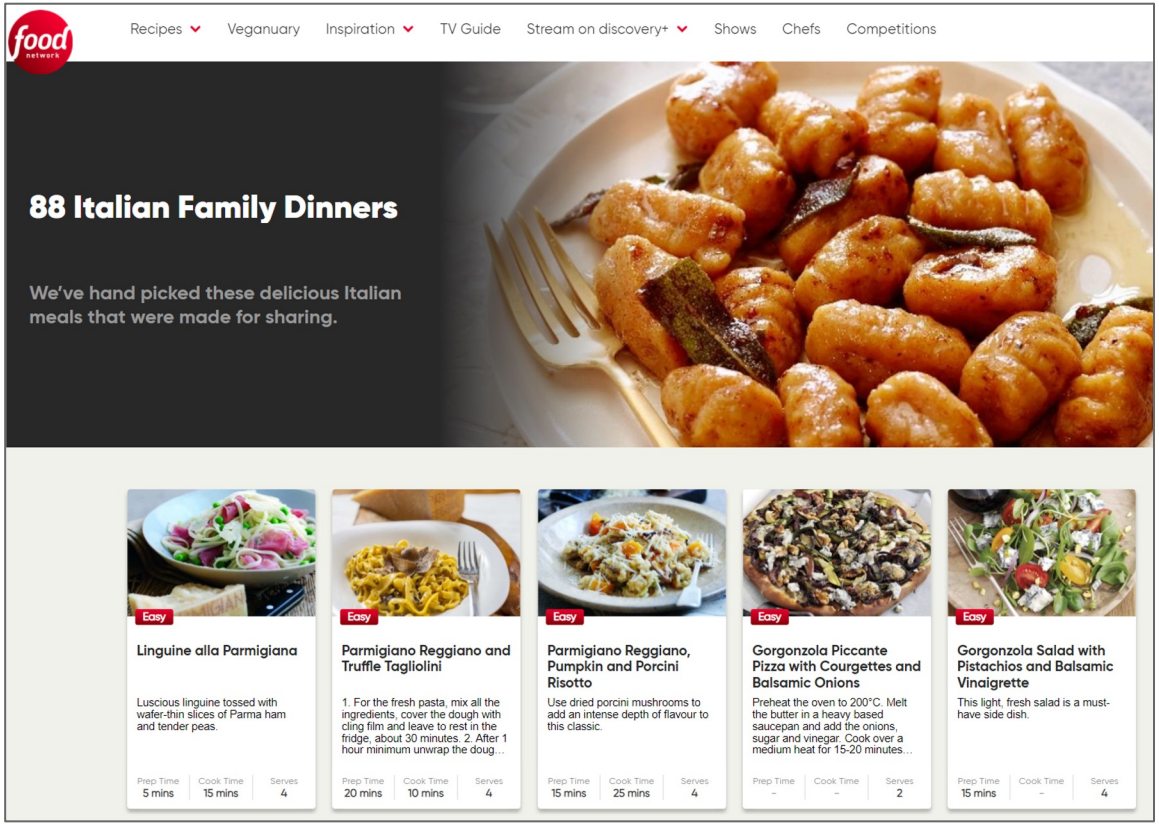
- Scrap all the links in the footer of **English** website of Vnexpress (https://e.vnexpress.net/):

- 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

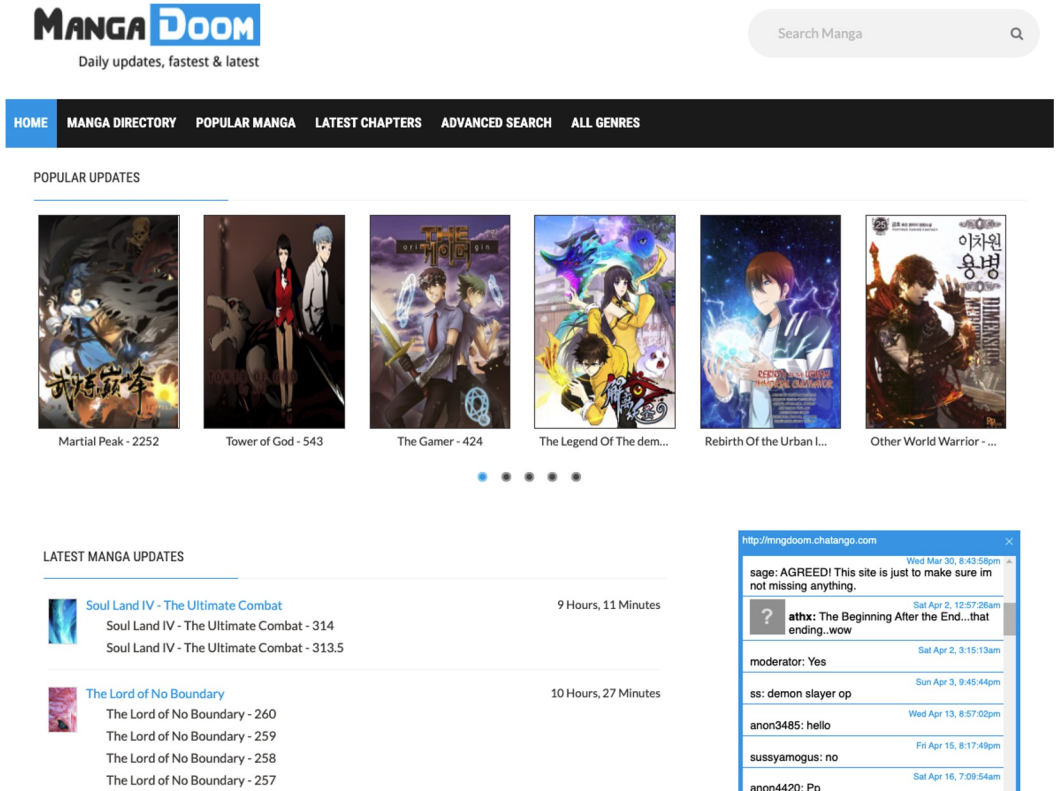
https://foodnetwork.co.uk/italian-family-dinners/



Scrap all Italian recipes, which should have:

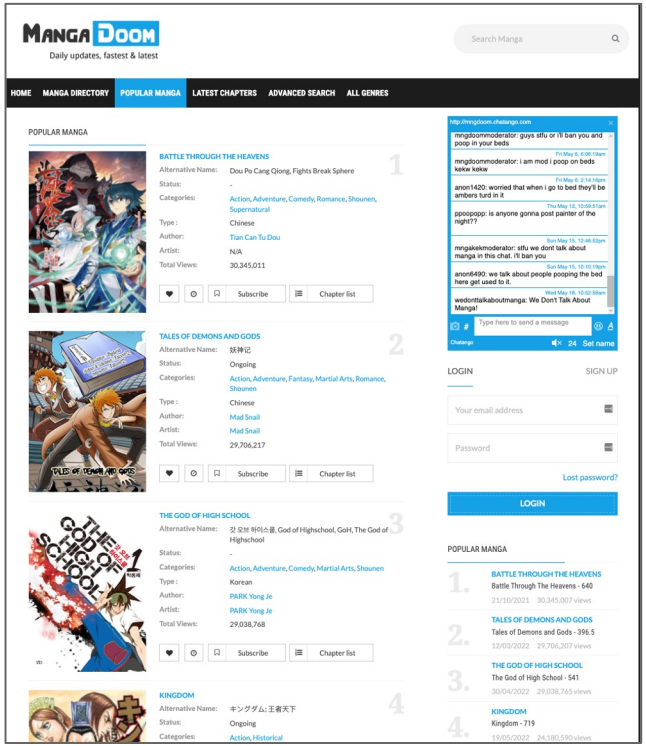
- Recipe title
- Short description
- Preparation time
- Cook time
- Number of serves

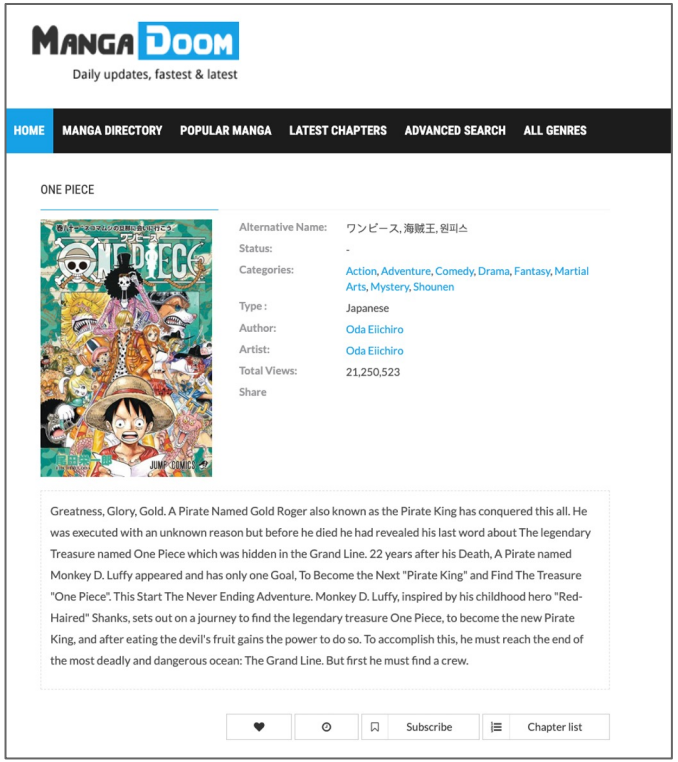
A Fun Project with Manga



Our Challenge

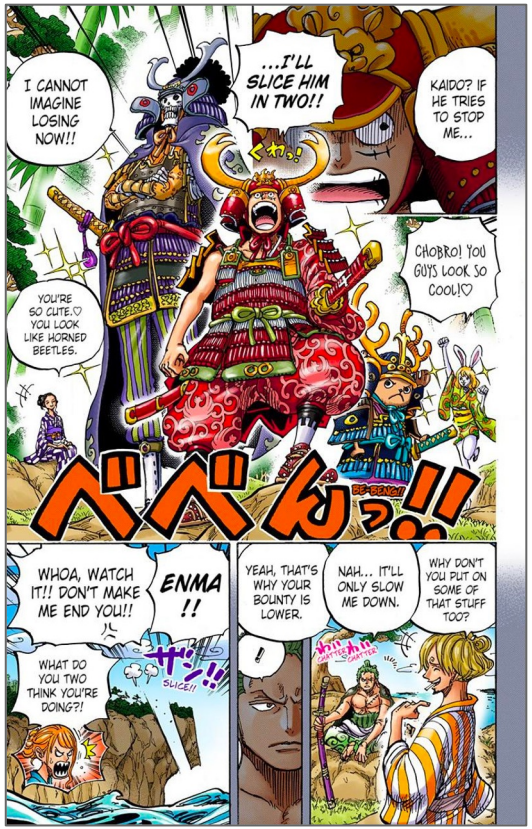
- Let’s scrap one of the best manga websites, “MangaDoom”!
  https://www.mngdoom.com/





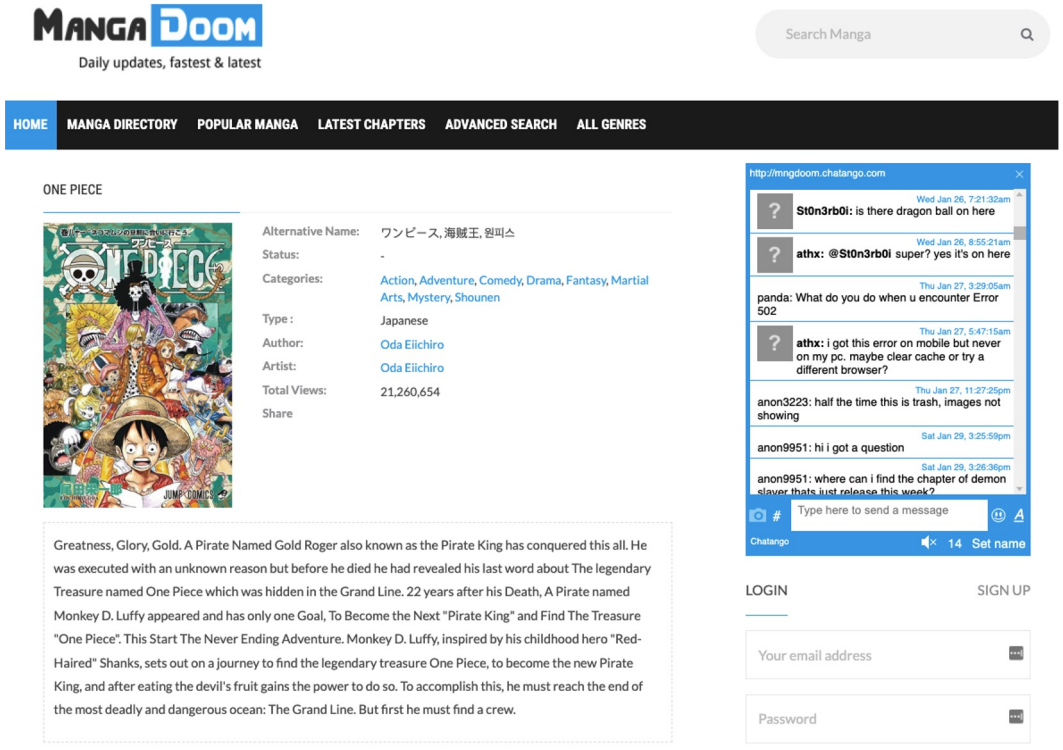
Let's pick a best manga to scrap them!

- "One Piece" is my favourite manga of all time! Let's scrap the whole manga!

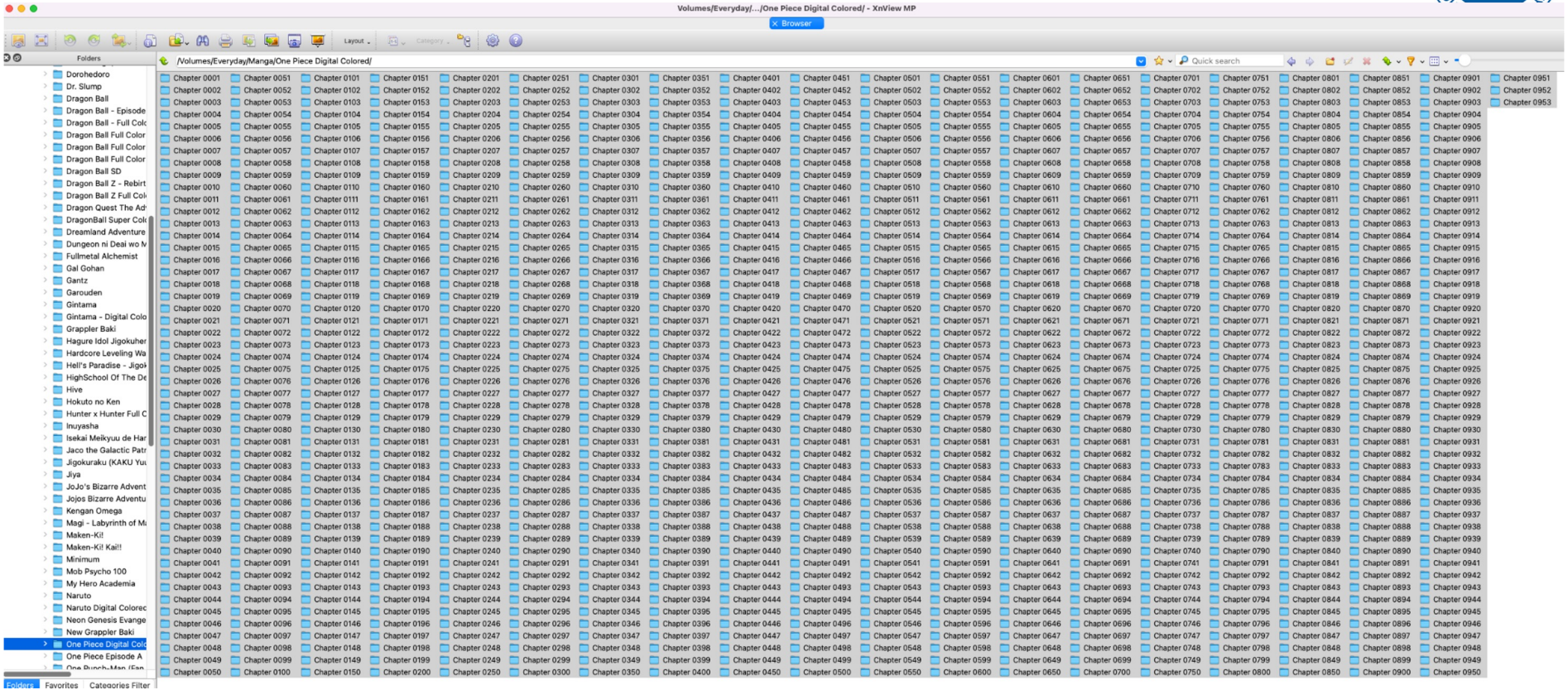


Let's do it together!

Scrap our favorite manga!



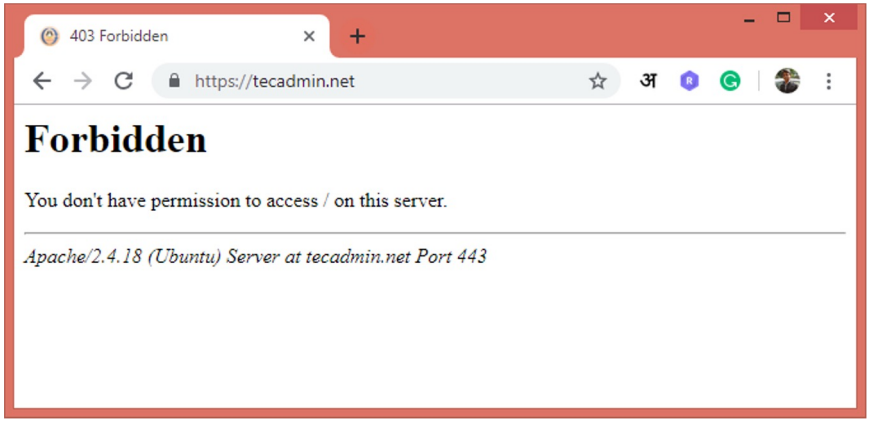
I'm a collector myself! Web Scraping is awesome!



Potential Problems & Solutions

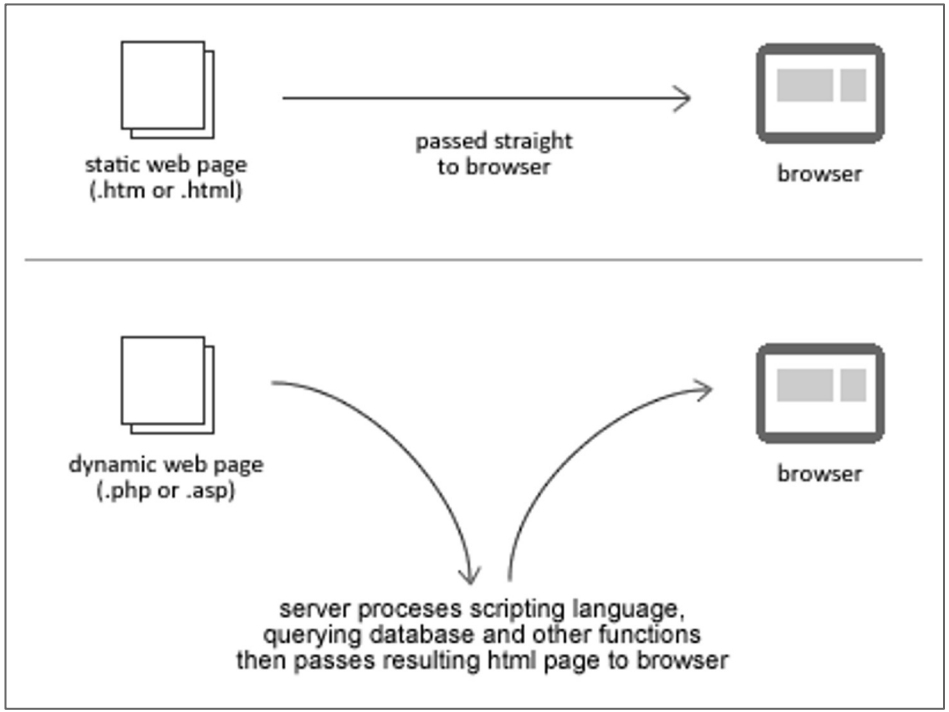
Potential Issues for Scraping

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

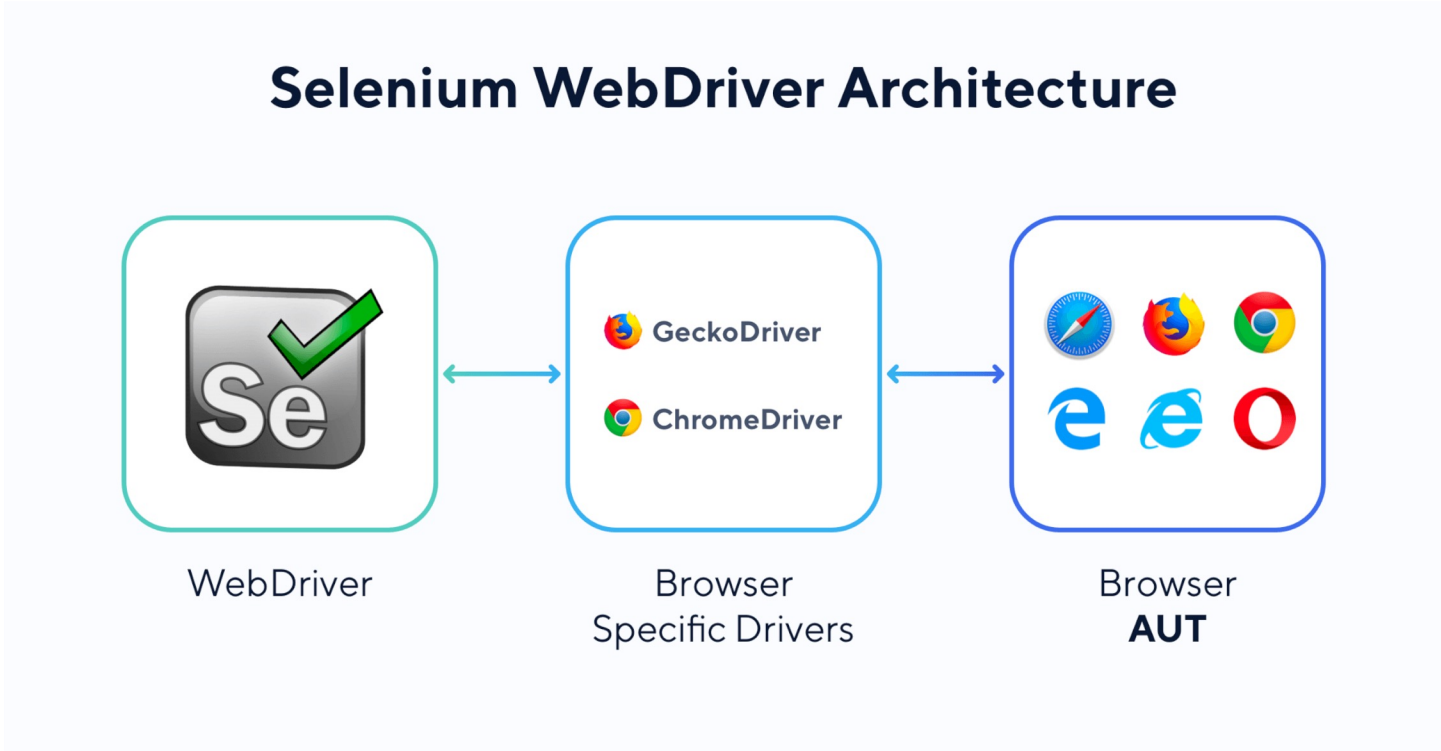


Static website vs Dynamic website

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



Selenium to scrape dynamic website



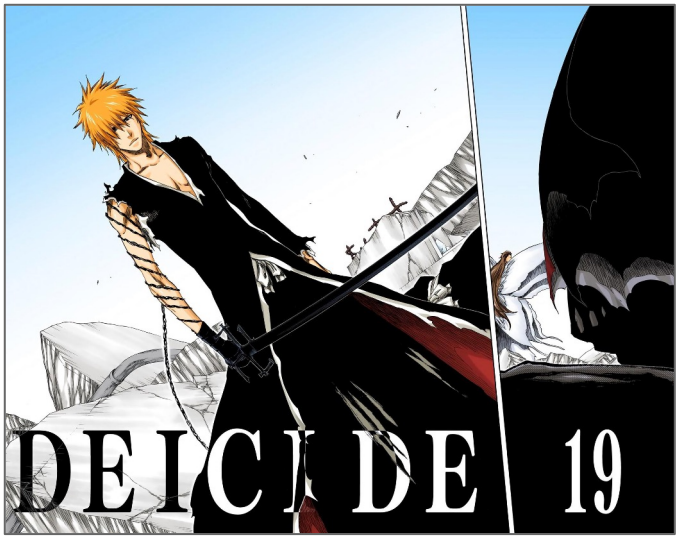
Fun Quiz

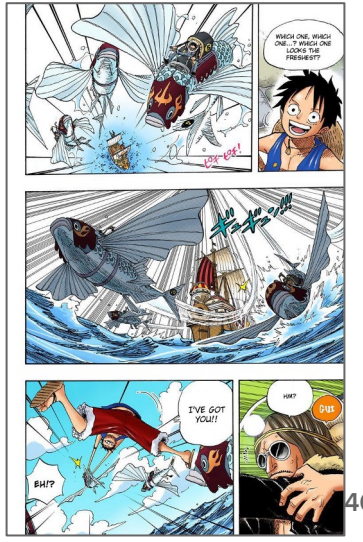
- Open popular ecommerce websites in Vietnam:
  - https://tiki.vn/
  - https://www.lazada.vn/
  - https://shopee.vn/
  - https://www.sendo.vn/
- Open your favourite website ever:
  - https://it.hcmiu.edu.vn/
- Open other websites:
  - https://baomoi.com/
  - https://mangasee123.com/
  - https://vi.wikipedia.org/wiki/Tr%C6%B0%E1%BB%9Dng_%C4%90%E1%BA%A1i_h%E1%BB%8Dc_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

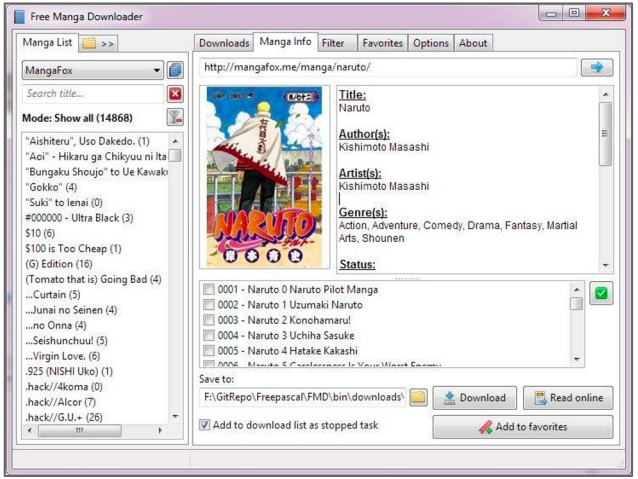
- **MangaSee** is **the only website** which got the **FULL-COLORED** chapters of big mangas like One Piece and Bleach!!!
- [Task] Can you figure out how to scrap either full-colored One Piece or Bleach?
  - https://mangasee123.com/manga/One-Piece-Digital-Colored-Comics
  - https://mangasee123.com/manga/Bleach-Color

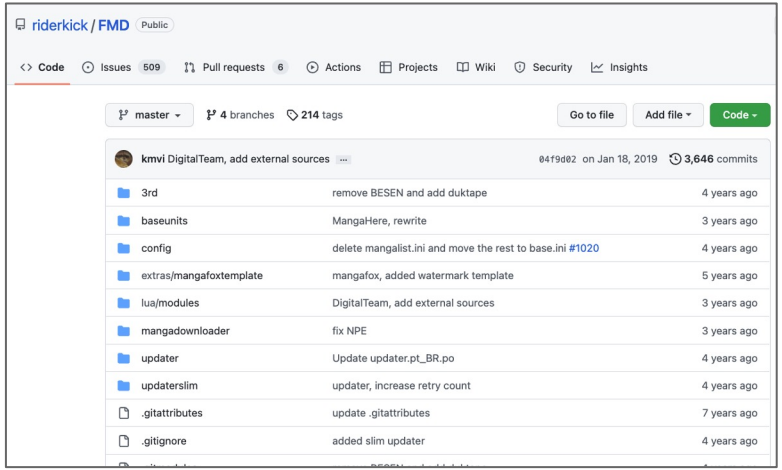


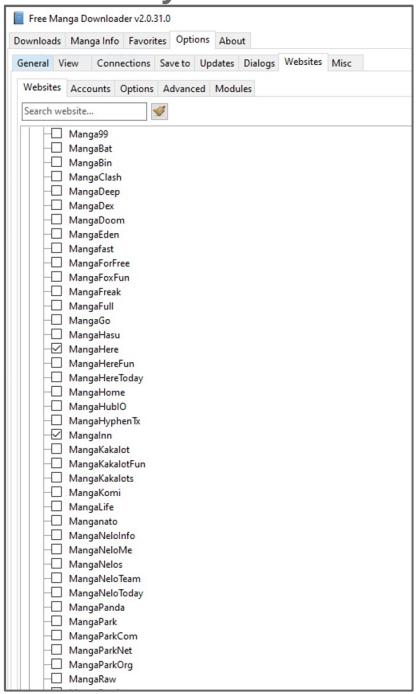


One of The Best Manga Downloader Project

- 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!







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