1. **What is web scraping?**

a) A method to download web pages

b) Extracting information from websites

c) Generating random data from the internet

d) Building websites from scratch

1. **Which package in R is commonly used for web scraping?**

a) ggplot2

b) dplyr

c) rvest

d) tidyr

1. **What function in the rvest package is used to extract data from HTML nodes?**

a) webscrape()

b) html\_element()

c) scrape\_html()

d) extract\_html()

1. **Which of the following is NOT a step in web scraping with R?**

a) Fetching HTML content from a website

b) Parsing HTML content

c) Training machine learning models

d) Extracting desired information

1. **What does the CSS selector ".title" select in an HTML document?**

a) All elements with the class "title"

b) The title tag

c) All elements with the ID "title"

d) All paragraphs with the class "title"

1. **Which function in R is used to download the HTML content of a webpage?** a)

a) download\_html()

b) get\_html()

c) read\_html()

d)fetch\_html()

1. **What is the purpose of the html\_text() function in rvest?**

a) Retrieves the text content of HTML nodes

b) Inserts text into HTML nodes

c) Creates HTML nodes from text

d) Deletes text from HTML nodes

1. **What is one common challenge in web scraping?**

a) Excessive availability of data

b) Websites consistently using the same structure

c) Websites blocking scrapers

d) Lack of internet connection

1. **Which of the following is a legal consideration in web scraping?**

a) Scraping any website without permission

b) Scraping copyrighted content without attribution

c) Scraping only public data

d) Scraping data for commercial purposes without restrictions

1. **What does the term "robots.txt" refer to in the context of web scraping?**

a) A file containing instructions for web crawlers

b) A type of HTML tag used for formatting text

c) A function in R for parsing URLs

d) A protocol for secure web communication

Note: In web scraping, "robots.txt" refers to a text file that website owners can create to communicate with web robots, also known as crawlers or spiders, about which parts of their site they can access and scrape. This file is typically located in the root directory of a website and is accessible to anyone who visits the site.

**Answers:**

1. b) Extracting information from websites
2. c) rvest
3. b) html\_element()
4. c) Training machine learning models
5. a) All elements with the class "title"
6. c) read\_html()
7. a) Retrieves the text content of HTML nodes
8. c) Websites blocking scrapers
9. c) Scraping only public data
10. a) A file containing instructions for web crawlers

**PART 2:** (50 points)

Question: Consider the HTML code provided below:

**# HTML code**

**html <- minimal\_html("**

**<title>Welcome to My Website</title>**

**<h1 id='main-heading'>Introduction</h1>**

**<p>This is the first paragraph.</p>**

**<h2>Main Content</h2>**

**<p class='content'>This is the main content of the webpage.</p>**

**<ul>**

**<li>List item 1</li>**

**<li>List item 2</li>**

**<li>List item 3</li>**

**</ul>**

**<h3>Conclusion</h3>**

**<p>Thanks for visiting!</p>**

**")**

Using the **html\_element()** and **html\_elements()** functions in R, write code to accomplish the following tasks:

* 1. **Write a code that returns.**

"" "Welcome to My Website"

* 1. **Write a code that returns.**

"This is the main content of the webpage."

* 1. **Write a code that returns.**

"List item 1\nList item 2\nList item 3"

* 1. **Write a code that returns.**

"List item 1"

* 1. **Write a code that returns.**

"List item 1" "List item 2" "List item 3"

**Solution:**

html\_text2(title\_elements)

[1] "" "Welcome to My Website"

> # Select the paragraph element with class 'content'

> paragraph\_content <- html\_element(html, ".content")

> html\_text2(paragraph\_content)

[1] "This is the main content of the webpage."

> # Select all list items within the unordered list

> u\_list <- html\_element(html, "ul")

> html\_text2(u\_list)

[1] "List item 1\nList item 2\nList item 3"

> # Select all list items within the ordered list

> l\_list <- html\_element(html, "li")

> html\_text2(l\_list)

[1] "List item 1"

> # Select all list items within the ordered list

> l\_list <- html\_elements(html, "li")

> html\_text2(l\_list)

[1] "List item 1" "List item 2" "List item 3"