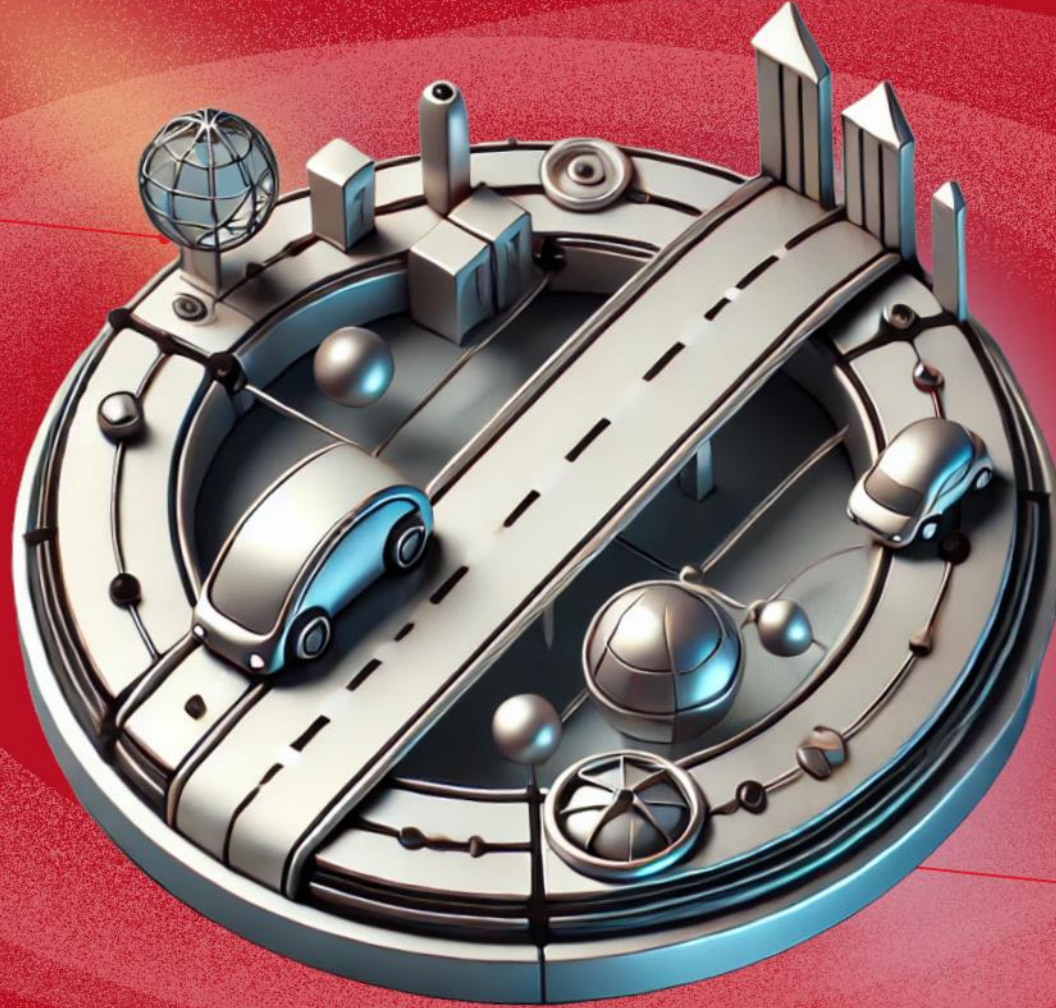


The Art and Science of Transportation Research in the AI Era

Web Scraping

Meng Cai



Learning goals



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

- #1** Understand the structure and components of HTML web pages
- #2** Become familiar with Python library “requests” and “BeautifulSoup”
- #3** Be able to scrape static HTML web pages
- #4** Have legal and ethical considerations in mind while scraping

Agenda



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

#1

Web page basics

#2

Web scraping with Python

Agenda



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

#1

Web page basics

#2

Web scraping with Python

#1.1 Components of a web page



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

HTML: Structure and content

CSS: Styling and presentation

JavaScript: Interactivity and dynamic content



HTML



HTML + CSS



HTML + CSS + JavaScript

Image source: <https://www.dreamstime.com/html-css-javascript-suit-as-explained-coding-layers-outline-diagram-website-project-development-stages-basic-skeletal-image233593998>

#1.2 HTML



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

HTML = HyperText Markup Language

Text with
embedded
links to other
text

A system using
tags to structure
and format text

```
1 <!DOCTYPE html>
2 <html lang="de">
3 <head>
4 <title>Home – TU Darmstadt</title>
5 <meta charset="UTF-8"/>
6 <meta name="viewport" content="width=device-width, initial-scale=1"/>
7 <meta http-equiv="X-UA-Compatible" content="ie=edge"/>
8 <meta name="dcterms.language" content="de"/>
9 <meta name="dcterms.title" content="Home"/>
10 <meta property="og:title" content="Home"/>
11 <meta property="og:site_name" content="TU Darmstadt"/>
12 <meta property="og:type" content="article"/>
13 <meta name="twitter:card" content="summary"/>
14 <meta name="twitter:title" content="Home"/>
15 <meta property="og:image" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_univer
16 <meta property="og:image:type" content="image/jpg"/>
17 <meta property="og:image:width" content="1180"/>
18 <meta property="og:image:height" content="664"/>
19 <meta property="og:image:alt" content=""/>
20 <meta name="twitter:image" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_univer
21 <meta name="twitter:image:alt" content=""/>
22 <meta name="ZOOMIMAGE" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_universit
23 <meta name="description" content="Homepage der Technischen Universität Darmstadt,
24 Universität in Hessen, Deutschland."/>
25 <meta property="og:description" content="Homepage der Technischen Universität Darmstadt,
26 Universität in Hessen, Deutschland."/>
27 <meta name="twitter:description" content="Homepage der Technischen Universität Darmstadt,
28 Universität in Hessen, Deutschland."/>
29 <meta name="dcterms.creator" content="Technische Universität Darmstadt"/>
30 <meta name="google-site-verification" content="MhVJ4YhH6BFMk3VMQ_0kf6dDKl64kAu05XQt9F__0qY"/>
31 <meta name="robots" content="index, follow"/>
32 <meta name="format-detection" content="telephone=no"/>
33 <meta name="generator" content="FirstSpirit 5.2.240809.5a859e4"/>
34 <link href="/media/resources/css_2/app_1.1.8.css" rel="stylesheet"/>
35 <link href="/media/resources/css_2/themes/css/theme-9c.css" rel="stylesheet"/>
36 <link rel="shortcut icon" type="image/x-icon" href="/media/resources/images_1/favicon_48x48.png">
37 <link rel="icon" type="image/png" href="/media/resources/images_1/favicon_48x48.png" sizes="48x48">
38 <link rel="icon" type="image/png" href="/media/resources/images_1/favicon_80x80.png" sizes="80x80">
39 <link rel="apple-touch-icon" sizes="80x80" href="/media/resources/images_1/favicon_80x80.png">
40 </head>
41 <body>
```

view-source:<https://www.tu-darmstadt.de/>

#1.3 HTML tags



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

<html>

</html>

```
1 <!DOCTYPE html>
2 <html lang="de">
3 <head>
4 <title>Home – TU Darmstadt</title>
5 <meta charset="UTF-8"/>
6 <meta name="viewport" content="width=device-width, initial-scale=1"/>
7 <meta http-equiv="X-UA-Compatible" content="ie=edge"/>
8 <meta name="dcterms.language" content="de"/>
9 <meta name="dcterms.title" content="Home"/>
10 <meta property="og:title" content="Home"/>
11 <meta property="og:site_name" content="TU Darmstadt"/>
12 <meta property="og:type" content="article"/>
13 <meta name="twitter:card" content="summary"/>
14 <meta name="twitter:title" content="Home"/>
15 <meta property="og:image" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_univer
16 <meta property="og:image:type" content="image/jpeg"/>
17 <meta property="og:image:width" content="1180"/>
18 <meta property="og:image:height" content="664"/>
19 <meta property="og:image:alt" content=""/>
20 <meta name="twitter:image" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_univer
21 <meta name="twitter:image:alt" content=""/>
22 <meta name="ZOOMIMAGE" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_universit
23 <meta name="description" content="Homepage der Technischen Universität Darmstadt,
24 Universität in Hessen, Deutschland."/>
25 <meta property="og:description" content="Homepage der Technischen Universität Darmstadt,
26 Universität in Hessen, Deutschland."/>
27 <meta name="twitter:description" content="Homepage der Technischen Universität Darmstadt,
28 Universität in Hessen, Deutschland."/>
29 <meta name="dcterms.creator" content="Technische Universität Darmstadt"/>
30 <meta name="google-site-verification" content="MhVJ4YhH6BFmk3VMQ_0kf6dDKl64kAu05XQt9F__0qY"/>
31 <meta name="robots" content="index, follow"/>
32 <meta name="format-detection" content="telephone=no"/>
33 <meta name="generator" content="FirstSpirit 5.2.240809.5a859e4"/>
34 <link href="/media/resources/css_2/app_1.1.8.css" rel="stylesheet"/>
35 <link href="/media/resources/css_2/themes_css/theme-9c.css" rel="stylesheet"/>
36 <link rel="shortcut icon" type="image/x-icon" href="/media/resources/images_1/favicon_48x48.png">
37 <link rel="icon" type="image/png" href="/media/resources/images_1/favicon_48x48.png" sizes="48x48">
38 <link rel="icon" type="image/png" href="/media/resources/images_1/favicon_80x80.png" sizes="80x80">
39 <link rel="apple-touch-icon" sizes="80x80" href="/media/resources/images_1/favicon_80x80.png">
40 </head>
41 <body>
```

#1.3 HTML tags



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

<html>

<head>

</head>

<body>

</body>

</html>

} contains metadata and
page title

} contains the main content
displayed on the page

```
1 <!DOCTYPE html>
2 <html lang="de">
3 <head>
4 <title>Home – TU Darmstadt</title>
5 <meta charset="UTF-8"/>
6 <meta name="viewport" content="width=device-width, initial-scale=1"/>
7 <meta http-equiv="X-UA-Compatible" content="ie=edge"/>
8 <meta name="dcterms.language" content="de"/>
9 <meta name="dcterms.title" content="Home"/>
10 <meta property="og:title" content="Home"/>
11 <meta property="og:site_name" content="TU Darmstadt"/>
12 <meta property="og:type" content="article"/>
13 <meta name="twitter:card" content="summary"/>
14 <meta name="twitter:title" content="Home"/>
15 <meta property="og:image" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_univer
16 <meta property="og:image:type" content="image/jpeg"/>
17 <meta property="og:image:width" content="1180"/>
18 <meta property="og:image:height" content="664"/>
19 <meta property="og:image:alt" content=""/>
20 <meta name="twitter:image" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_univer
21 <meta name="twitter:image:alt" content=""/>
22 <meta name="ZOOMIMAGE" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_universit
23 <meta name="description" content="Homepage der Technischen Universität Darmstadt,
24 Universität in Hessen, Deutschland."/>
25 <meta property="og:description" content="Homepage der Technischen Universität Darmstadt,
26 Universität in Hessen, Deutschland."/>
27 <meta name="twitter:description" content="Homepage der Technischen Universität Darmstadt,
28 Universität in Hessen, Deutschland."/>
29 <meta name="dcterms.creator" content="Technische Universität Darmstadt"/>
30 <meta name="google-site-verification" content="MhVJ4YhH6BFMk3VMQ_0kf6dDKL64kAu05XQt9F__0qY"/>
31 <meta name="robots" content="index,follow"/>
32 <meta name="format-detection" content="telephone=no"/>
33 <meta name="generator" content="FirstSpirit 5.2.240809.5a859e4"/>
34 <link href="/media/resources/css_2/app_1.1.8.css" rel="stylesheet"/>
35 <link href="/media/resources/css_2/themes/css/theme-9c.css" rel="stylesheet"/>
36 <link rel="shortcut icon" type="image/x-icon" href="/media/resources/images_1/favicon_48x48.png">
37 <link rel="icon" type="image/png" href="/media/resources/images_1/favicon_48x48.png" sizes="48x48">
38 <link rel="icon" type="image/png" href="/media/resources/images_1/favicon_80x80.png" sizes="80x80">
39 <link rel="apple-touch-icon" sizes="80x80" href="/media/resources/images_1/favicon_80x80.png">
40 </head>
41 <body>
```


#1.3 HTML tags



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

```
<html>
```

```
<head>
```

```
</head>
```

```
<body>
```

```
<p>Here's a paragraph of text!</p>
```

```
<p>Here's a second paragraph of text!</p>
```

} Text inside `<p>` tags is displayed as separate paragraphs

```
</body>
```

```
</html>
```

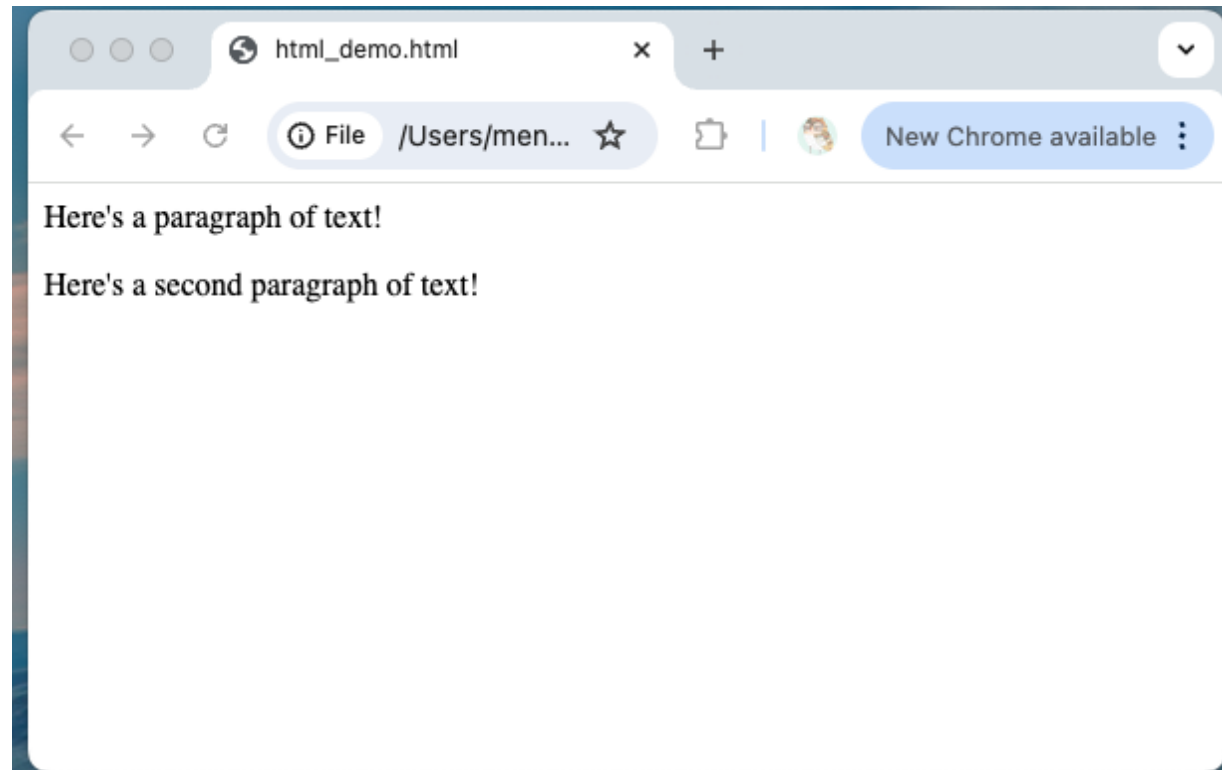
#1.3 HTML tags



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt



#1.3 HTML tags



Tags can be nested within other tags

- Parent: A tag that contains another tag
- Child: A tag that is contained within another tag
- Sibling: Tags that share the same parent

```
<html>
```

```
<head>
```

```
</head>
```

```
<body>
```

```
<p>Here's a paragraph of text!</p>
```

```
<p>Here's a second paragraph of text!</p>
```

```
</body>
```

```
</html>
```

#1.3 HTML tags



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

<p>

Here's a paragraph of text!

TUDa → defines a hyperlink

</p>

href attribute specifies the destination URL

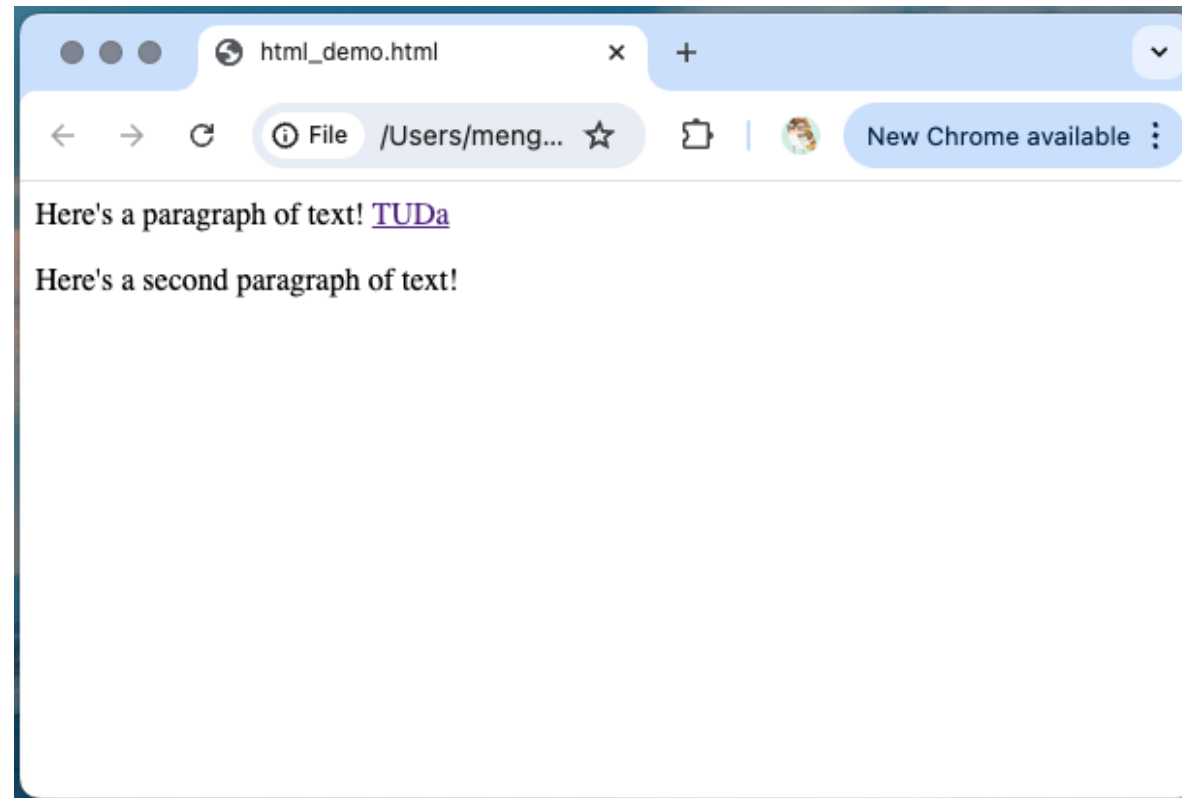
#1.3 HTML tags



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt



#1.3 HTML tags



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

Common HTML tags

- `<div>`: Defines a division or a section
- ``: Makes text bold
- `<i>`: Italicizes text
- `<table>`: Creates a table structure
- `<form>`: Defines an input form

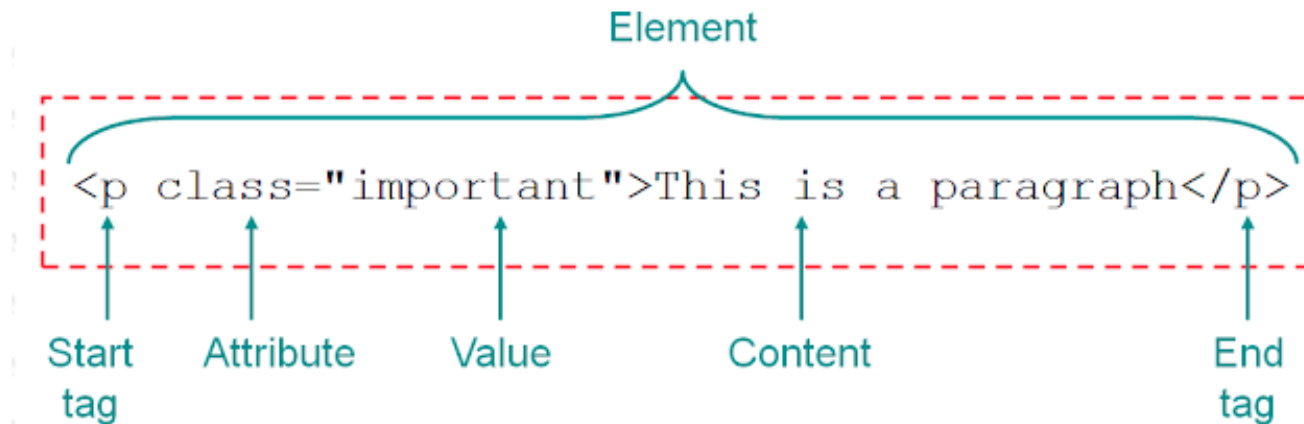
Many more tags exist for various purposes

#1.4 HTML attributes



All HTML elements can have attributes

- Attributes provide additional information about elements
- Common attributes include href, class, id, src, etc.
- Syntax: **<tag attribute="value">** e.g. ``



Img source: <https://www.onlinedesignteacher.com/2016/03/html-attributes.html>

#1.4 HTML attributes



Two of the most important attributes are class and id.

- **class:** Assigns one or more names to an element
Can be shared among multiple elements
- **id:** Assigns a unique identifier to an element
Should be unique within the entire HTML document

```
<p class="bold-paragraph">  
  Here's a paragraph of text!  
  <a href="https://www.tu-darmstadt.de" id="link">TUDa</a>  
</p>  
<p class="bold-paragraph extra-large">  
  Here's a second paragraph of text!  
  <a href="https://www.python.org" class="extra-large">Python</a>  
</p>
```

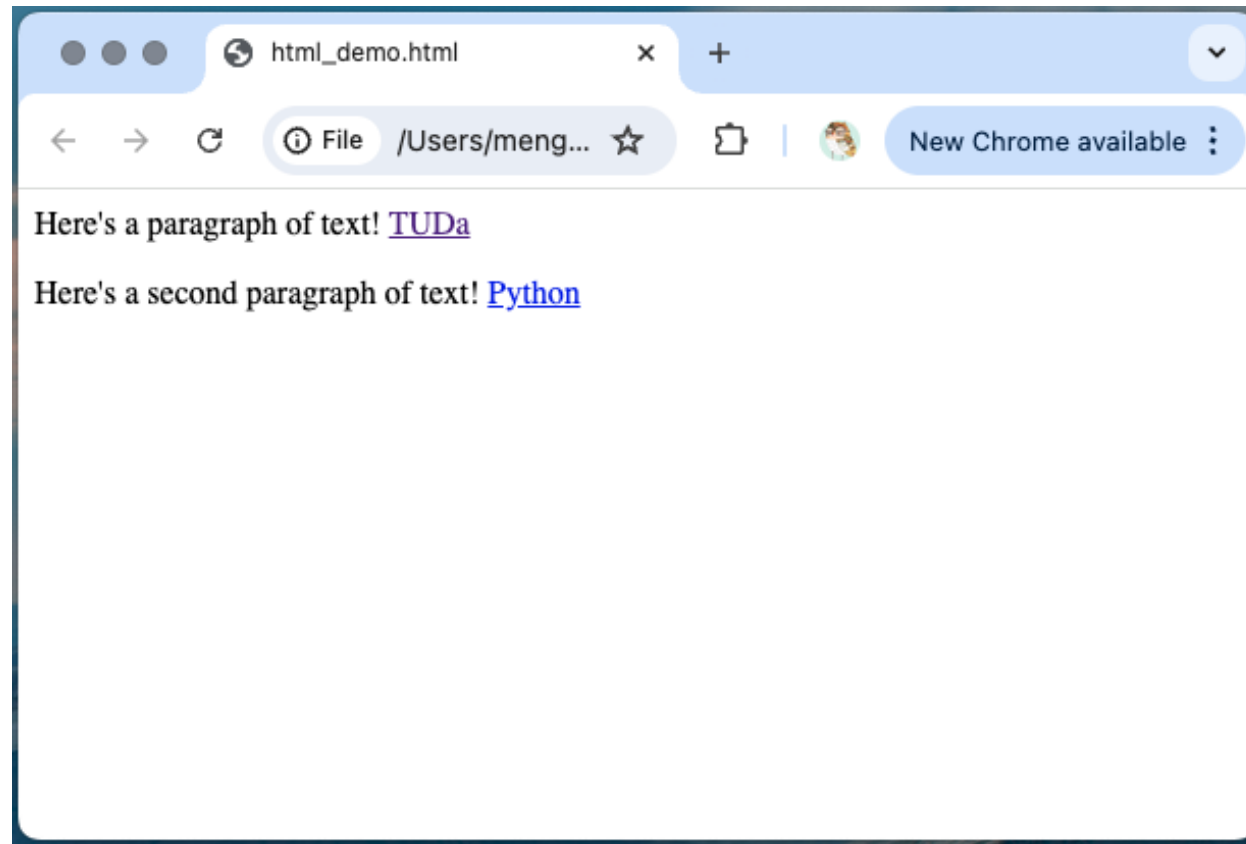

#1.4 HTML attributes



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt



#1.5 Activity: Create an HTML file



Open Jupyter Notebook and create a new Text File. Rename the untitled.txt as a_name.html.

An example

```
1 <!DOCTYPE html>
2 <html>
3   <head>
4   </head>
5   <body>
6     <p class="bold-paragraph">
7       Here's a paragraph of text!
8       <a href="https://www.tu-darmstadt.de" id="link">TUDa</a>
9     </p>
10    <p class="bold-paragraph extra-large">
11      Here's a second paragraph of text!
12      <a href="https://www.python.org" class="extra-large">Python</a>
13    </p>
14  </body>
15 </html>
```

#1.6 Summary of web page basics



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

- Web pages are built using HTML for structure and content
- HTML structures content with tags and attributes
- Tags can be nested to create parent-child relationships
- Common tags include `<html>`, `<head>`, `<body>`, `<p>`, and `<a>`
- Attributes like `class` and `id` help identify and select elements

Agenda



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

#1

Web page basics

#2

Web scraping with Python

#2.1 What is web scraping?



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

- Automated extraction of data from websites
- Converts unstructured data into structured data
- Involves fetching and parsing web content

#2.2 Why is web scraping useful?



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

Web scraping is particularly useful in scenarios where data is needed but not readily available in structured formats.

- Volume: Efficient data collection, especially in large amount
- Time: Access to real-time data updates
- Scale: Enables large-scale data collection and analysis
- Diversity: Access to a wide range of topics and languages

Web scraping is a primary method of collecting text data for Large Language Models.

#2.3 Common challenges



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

- Websites using JavaScript to load content dynamically
- IP Blocking and CAPTCHAs
- Changing Website Structures

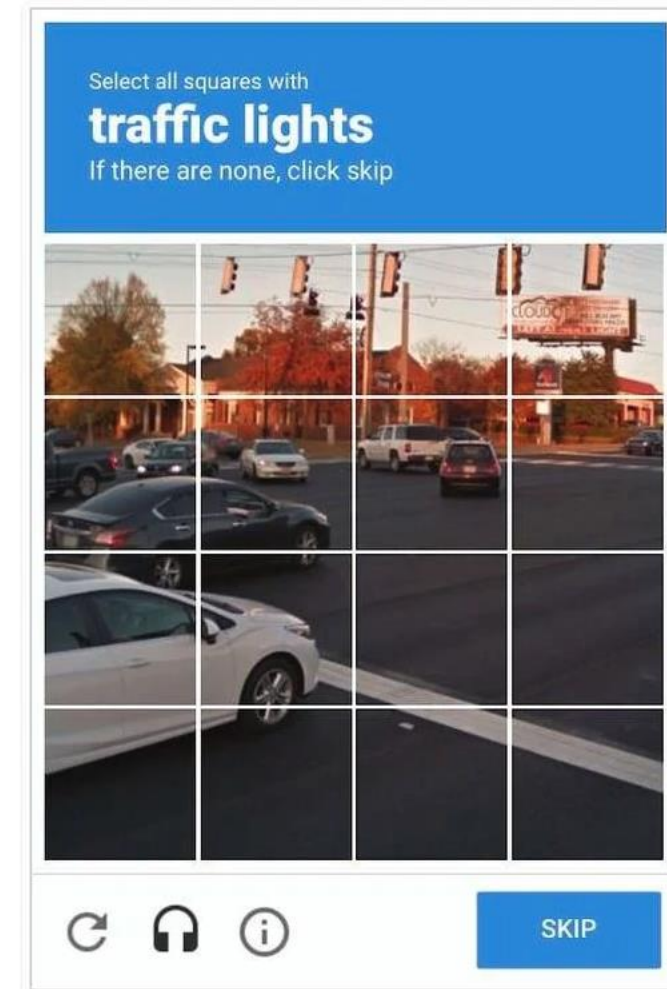


Image source: <https://www.zenrows.com/blog/avoid-captcha#types-of-captcha>

#2.4 Legal and ethical considerations



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

403

Forbidden

Access to this resource on the server is denied!

Legal

- Terms of Service
- Robots.txt Protocol
- Data Protection and Privacy Laws
- ...

Ethical

- Respecting ownership of data
- Avoiding data misuse
- Avoid overloading servers

Image source: <https://www.crazydomains.com/help/article/403-forbidden-error-explained>

#2.5 Activity: learn more about robots.txt



```
# robots.txt
#
# This file is to prevent the crawling and indexing of certain parts
# of your site by web crawlers and spiders run by sites like Yahoo!
# and Google. By telling these "robots" where not to go on your site,
# you save bandwidth and server resources.
#
# This file will be ignored unless it is at the root of your host:
# Used:    http://example.com/robots.txt
# Ignored: http://example.com/site/robots.txt
#
# For more information about the robots.txt standard, see:
# http://www.robotstxt.org/robotstxt.html
```

```
User-agent: *
# CSS, JS, Images
Allow: /core/*.css$
Allow: /core/*.css?
Allow: /core/*.js$
Allow: /core/*.js?
Allow: /core/*.gif
Allow: /core/*.jpg
Allow: /core/*.jpeg
Allow: /core/*.png
Allow: /core/*.svg
Allow: /profiles/*.css$
Allow: /profiles/*.css?
Allow: /profiles/*.js$
Allow: /profiles/*.js?
Allow: /profiles/*.gif
Allow: /profiles/*.jpg
Allow: /profiles/*.jpeg
Allow: /profiles/*.png
Allow: /profiles/*.svg
# Directories
Disallow: /core/
Disallow: /profiles/
# Files
Disallow: /README.txt
Disallow: /web.config
# Paths (clean URLs)
Disallow: /admin/
Disallow: /comment/reply/
Disallow: /filter/tips
Disallow: /node/add/
Disallow: /search/
Disallow: /user/register/
Disallow: /user/password/
Disallow: /user/login/
Disallow: /user/logout/
.._..
```

robots.txt is a voluntary standard for well-behaved scrapers.

1. Please go to <http://www.robotstxt.org>

2. Skim through “About /robots.txt”

3. Find out the path to robots.txt and try it out with TUDa (www.tu-darmstadt.de) or GitHub (www.github.com/)

#2.6 Activity: let's scrape a page



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

Please go to the Jupyter notebook.

#2.7 Summary of web scraping with Python



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

- Web scraping is an automated process of extracting data from websites
- Transforms unstructured HTML content into structured data.
- Basic Steps in web scraping:
 - Send an HTTP request
 - Parse the HTML content
 - Locate the HTML elements containing the desired data
 - Extract data
- Make sure to scrape legally and ethically

Learning goals



TECHNISCHE
UNIVERSITÄT
DARMSTADT



Institut für
Verkehrsplanung
und Verkehrstechnik
TU Darmstadt

- #1** Understand the structure and components of HTML web pages
- #2** Become familiar with Python library “requests” and “BeautifulSoup”
- #3** Be able to scrape static HTML web pages
- #4** Have legal and ethical considerations in mind while scraping