The Art and Science of Transportation Research in the AI Era

Web Scraping

Meng Cai

Learning goals

#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

#1 Web page basics

#2 Web scraping with Python

Agenda

#1 Web page basics

#2 Web scraping with Python

#1.1 Components of a web page

HTML: Structure and content

CSS: Styling and presentation

JavaScript: Interactivity and dynamic content

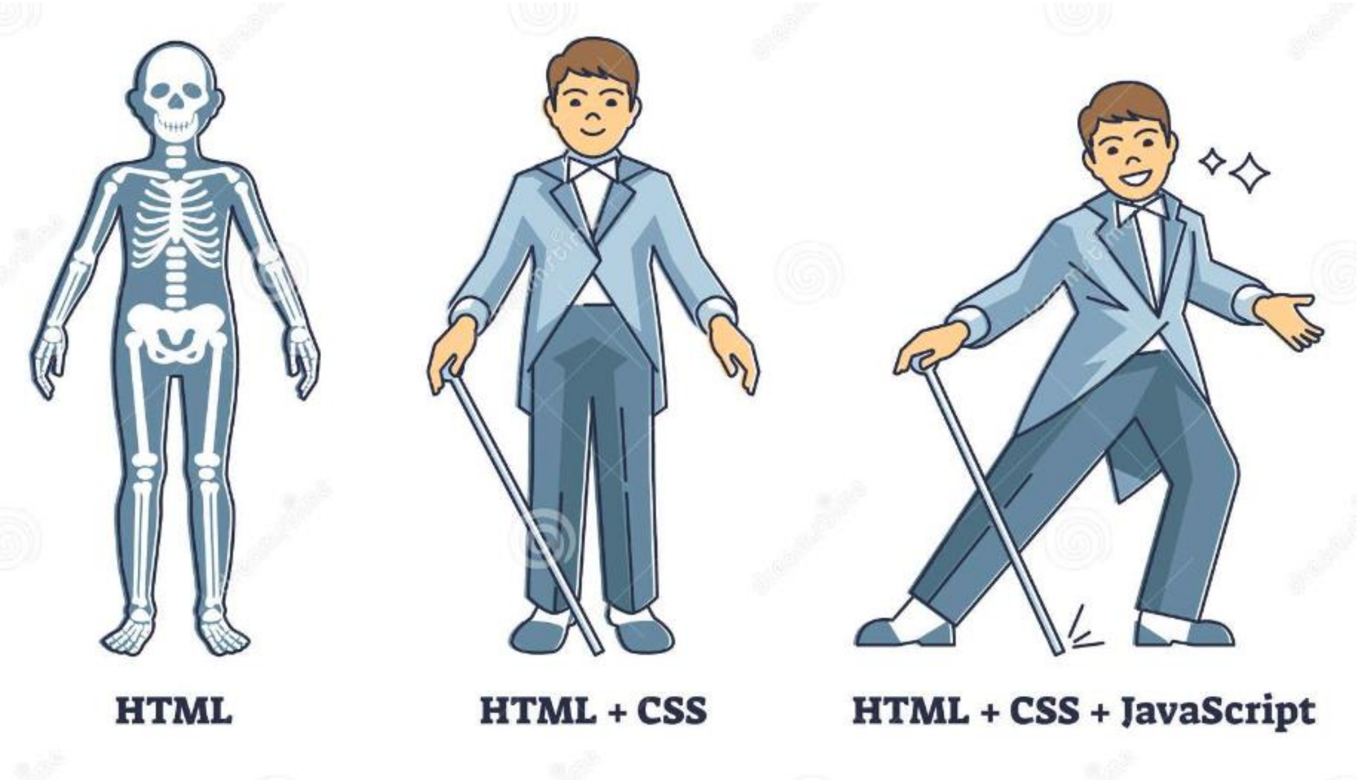
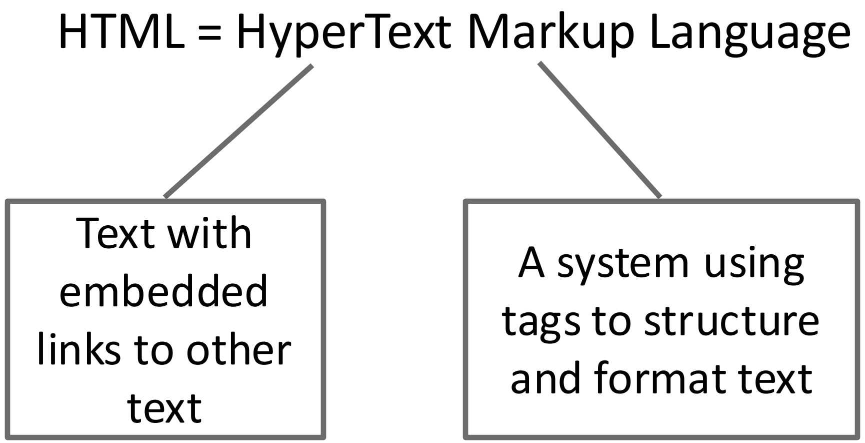


Image source: https://www.dreamstime.com/html-css-javascript-suit-as-explained-coding-layers-outline-diagram-website-project-development-stages-basic-skeletal-image233593998

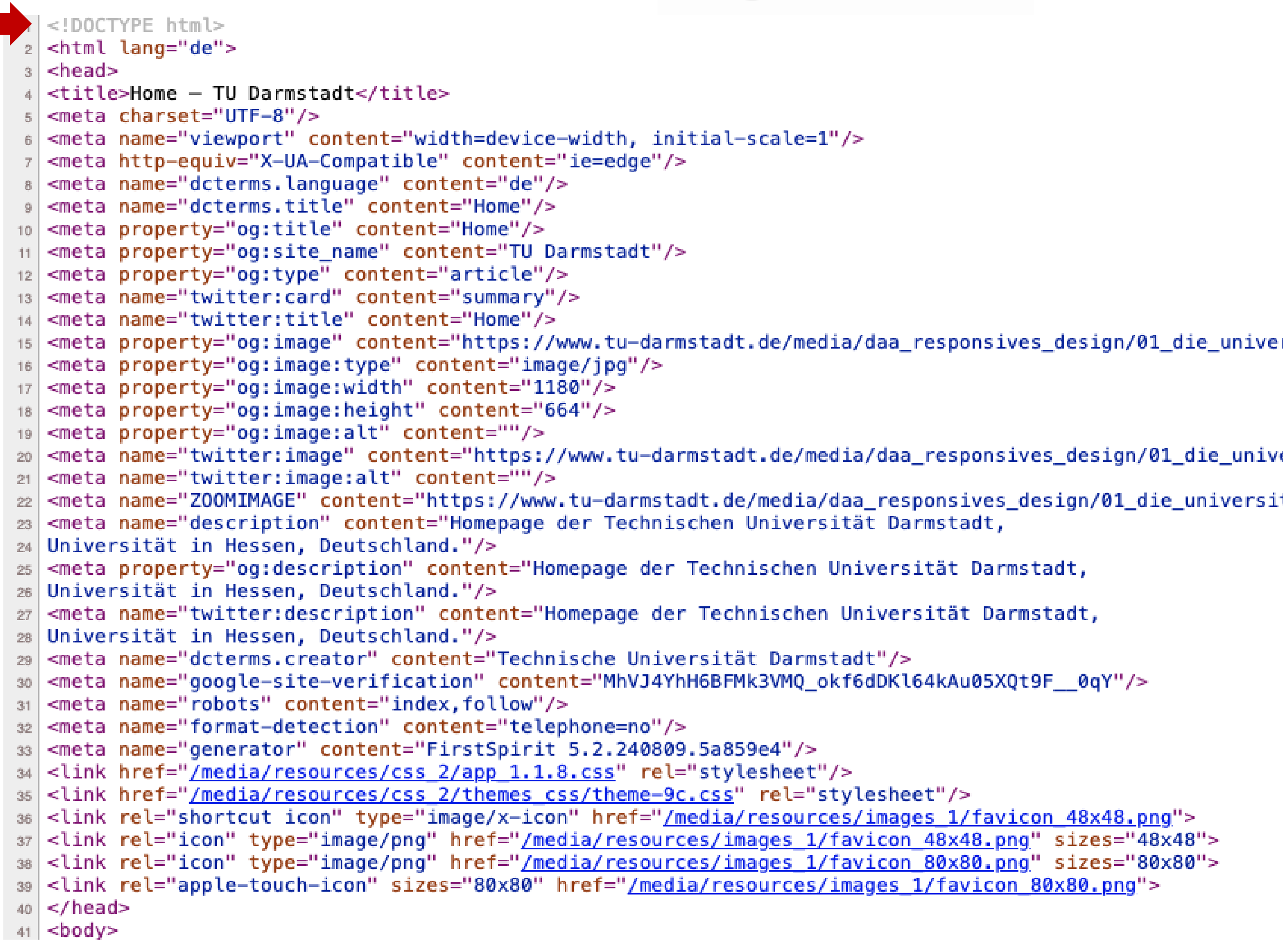
#1.2 HTML



HTML = HyperText Markup Language

Text with embedded links to other text

A system using tags to structure and format text

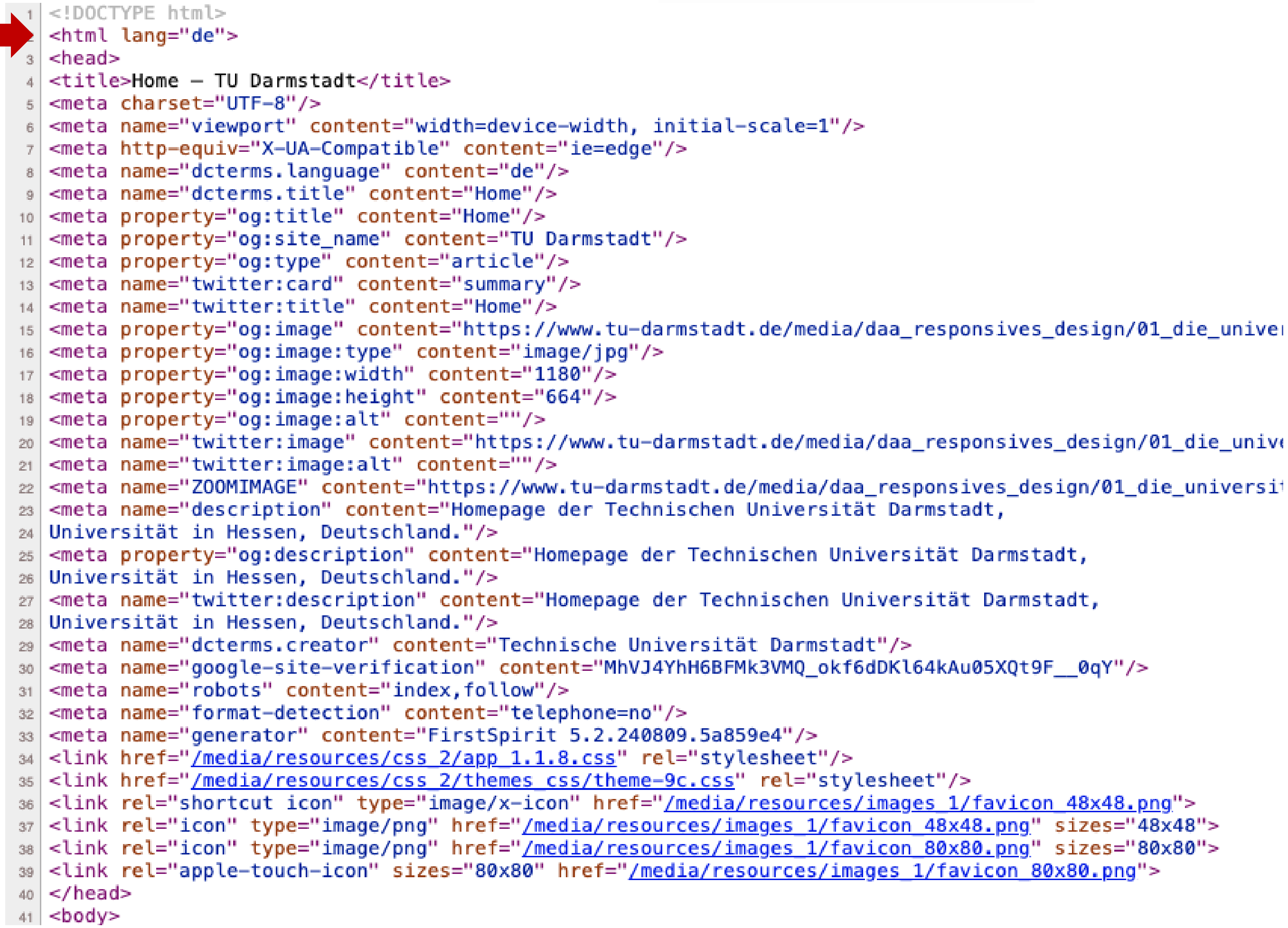


```
<!DOCTYPE html>
<html lang="de">
<head>
<title>Home – TU Darmstadt</title>
<meta charset="UTF-8"/>
<meta name="viewport" content="width=device-width, initial-scale=1"/>
<meta http-equiv="X-UA-Compatible" content="ie=edge"/>
<meta name="dcterms.language" content="de"/>
<meta name="dcterms.title" content="Home"/>
<meta property="og:title" content="Home"/>
<meta property="og:site_name" content="TU Darmstadt"/>
<meta property="og:type" content="article"/>
<meta name="twitter:card" content="summary"/>
<meta name="twitter:title" content="Home"/>
<meta property="og:image" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_unive
<meta property="og:image:type" content="image/jpg"/>
<meta property="og:image:width" content="1180"/>
<meta property="og:image:height" content="664"/>
<meta property="og:image:alt" content=""/>
<meta name="twitter:image" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_univ
<meta name="twitter:image:alt" content=""/>
<meta name="ZOOMIMAGE" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_universi
<meta name="description" content="Homepage der Technischen Universität Darmstadt,
Universität in Hessen, Deutschland."/>
<meta property="og:description" content="Homepage der Technischen Universität Darmstadt,
Universität in Hessen, Deutschland."/>
<meta name="twitter:description" content="Homepage der Technischen Universität Darmstadt,
Universität in Hessen, Deutschland."/>
<meta name="dcterms.creator" content="Technische Universität Darmstadt"/>
<meta name="google-site-verification" content="MhVJ4YhH6BFMk3VMQ_okf6dDKl64kAu05XQt9F__0qY"/>
<meta name="robots" content="index,follow"/>
<meta name="format-detection" content="telephone=no"/>
<meta name="generator" content="FirstSpirit 5.2.240809.5a859e4"/>
<link href="/media/resources/css_2/app_1.1.8.css" rel="stylesheet"/>
<link href="/media/resources/css_2/themes_css/theme-9c.css" rel="stylesheet"/>
<link rel="shortcut icon" type="image/x-icon" href="/media/resources/images_1/favicon_48x48.png">
<link rel="icon" type="image/png" href="/media/resources/images_1/favicon_48x48.png" sizes="48x48">
<link rel="icon" type="image/png" href="/media/resources/images_1/favicon_80x80.png" sizes="80x80">
<link rel="apple-touch-icon" sizes="80x80" href="/media/resources/images_1/favicon_80x80.png">
</head>
<body>
```

Type in Chrome: “view-source:https://www.tu-darmstadt.de/”

#1.3 HTML tags

```
<html>

</html>
```



```
<!DOCTYPE html>
<html lang="de">
<head>
<title>Home – TU Darmstadt</title>
<meta charset="UTF-8"/>
<meta name="viewport" content="width=device-width, initial-scale=1"/>
<meta http-equiv="X-UA-Compatible" content="ie=edge"/>
<meta name="dcterms.language" content="de"/>
<meta name="dcterms.title" content="Home"/>
<meta property="og:title" content="Home"/>
<meta property="og:site_name" content="TU Darmstadt"/>
<meta property="og:type" content="article"/>
<meta name="twitter:card" content="summary"/>
<meta name="twitter:title" content="Home"/>
<meta property="og:image" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_unive
<meta property="og:image:type" content="image/jpg"/>
<meta property="og:image:width" content="1180"/>
<meta property="og:image:height" content="664"/>
<meta property="og:image:alt" content=""/>
<meta name="twitter:image" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_unive
<meta name="twitter:image:alt" content=""/>
<meta name="ZOOMIMAGE" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_universi
<meta name="description" content="Homepage der Technischen Universität Darmstadt,
Universität in Hessen, Deutschland."/>
<meta property="og:description" content="Homepage der Technischen Universität Darmstadt,
Universität in Hessen, Deutschland."/>
<meta name="twitter:description" content="Homepage der Technischen Universität Darmstadt,
Universität in Hessen, Deutschland."/>
<meta name="dcterms.creator" content="Technische Universität Darmstadt"/>
<meta name="google-site-verification" content="MhVJ4YhH6BFMk3VMQ_okf6dDKl64kAu05XQt9F__0qY"/>
<meta name="robots" content="index,follow"/>
<meta name="format-detection" content="telephone=no"/>
<meta name="generator" content="FirstSpirit 5.2.240809.5a859e4"/>
<link href="/media/resources/css_2/app_1.1.8.css" rel="stylesheet"/>
<link href="/media/resources/css_2/themes_css/theme-9c.css" rel="stylesheet"/>
<link rel="shortcut icon" type="image/x-icon" href="/media/resources/images_1/favicon_48x48.png">
<link rel="icon" type="image/png" href="/media/resources/images_1/favicon_48x48.png" sizes="48x48">
<link rel="icon" type="image/png" href="/media/resources/images_1/favicon_80x80.png" sizes="80x80">
<link rel="apple-touch-icon" sizes="80x80" href="/media/resources/images_1/favicon_80x80.png">
</head>
<body>
```

#1.3 HTML tags

<html>

<head>

contains metadata and page title

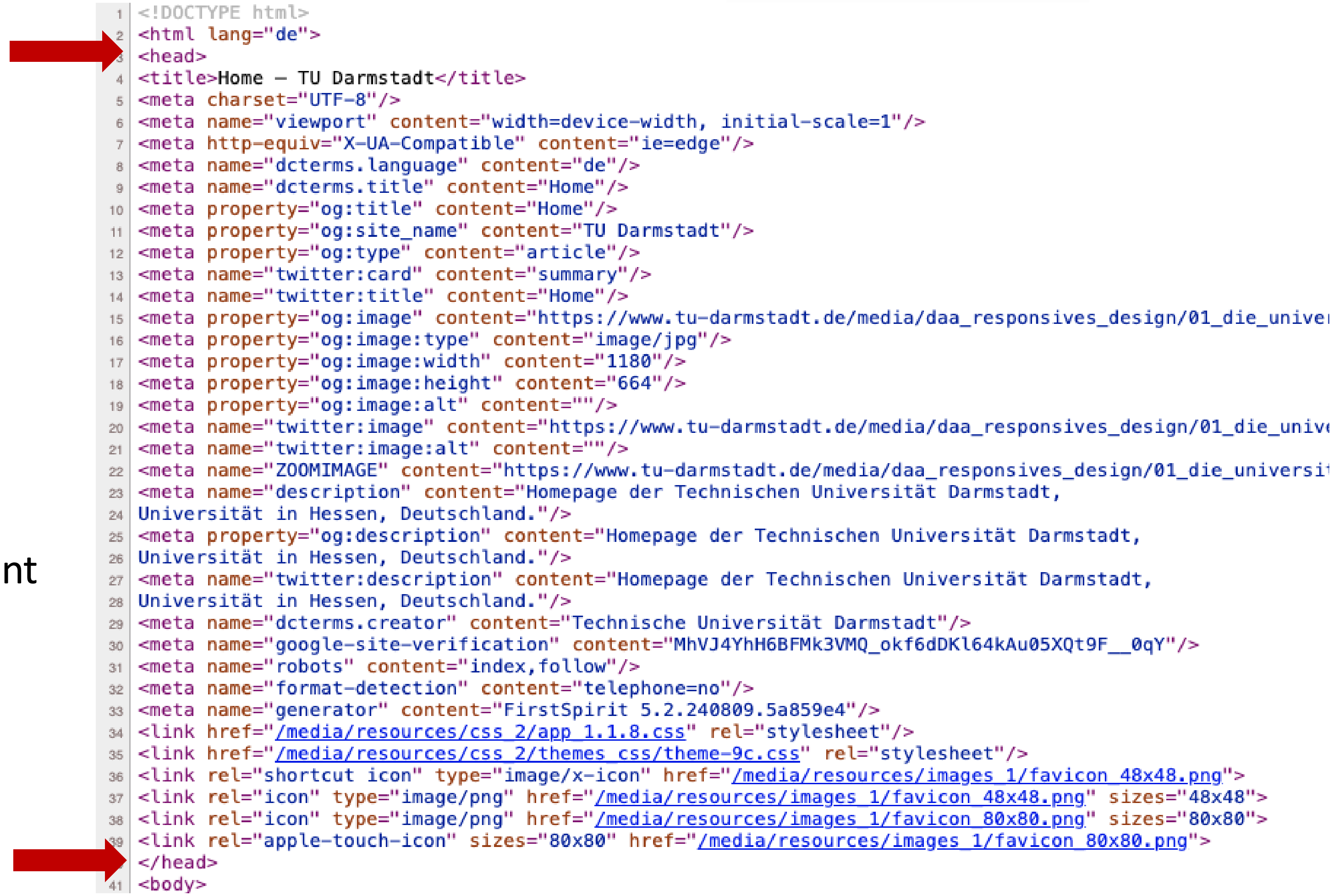
</head>

<body>

contains the main content displayed on the page

</body>

</html>



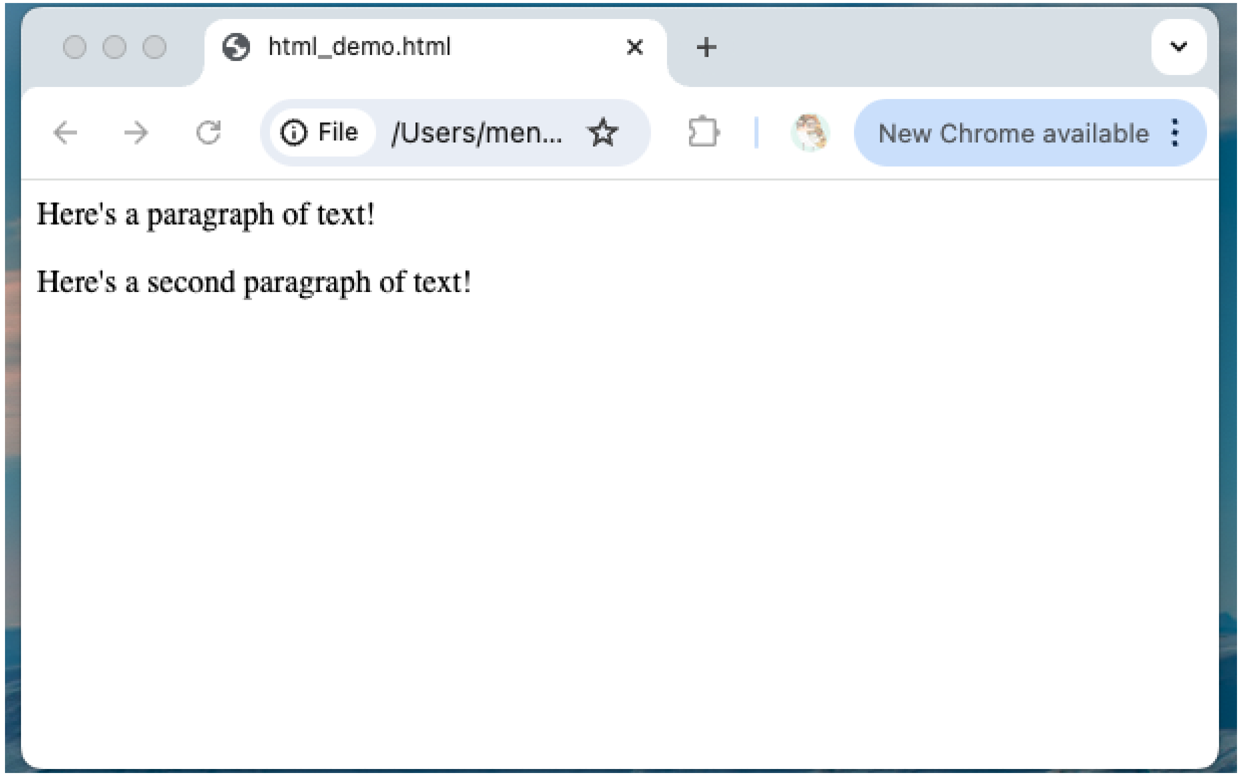
```
<!DOCTYPE html>
<html lang="de">
<head>
<title>Home – TU Darmstadt</title>
<meta charset="UTF-8"/>
<meta name="viewport" content="width=device-width, initial-scale=1"/>
<meta http-equiv="X-UA-Compatible" content="ie=edge"/>
<meta name="dcterms.language" content="de"/>
<meta name="dcterms.title" content="Home"/>
<meta property="og:title" content="Home"/>
<meta property="og:site_name" content="TU Darmstadt"/>
<meta property="og:type" content="article"/>
<meta name="twitter:card" content="summary"/>
<meta name="twitter:title" content="Home"/>
<meta property="og:image" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_unive
<meta property="og:image:type" content="image/jpg"/>
<meta property="og:image:width" content="1180"/>
<meta property="og:image:height" content="664"/>
<meta property="og:image:alt" content=""/>
<meta name="twitter:image" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_univ
<meta name="twitter:image:alt" content=""/>
<meta name="ZOOMIMAGE" content="https://www.tu-darmstadt.de/media/daa_responsives_design/01_die_universi
<meta name="description" content="Homepage der Technischen Universität Darmstadt,
Universität in Hessen, Deutschland."/>
<meta property="og:description" content="Homepage der Technischen Universität Darmstadt,
Universität in Hessen, Deutschland."/>
<meta name="twitter:description" content="Homepage der Technischen Universität Darmstadt,
Universität in Hessen, Deutschland."/>
<meta name="dcterms.creator" content="Technische Universität Darmstadt"/>
<meta name="google-site-verification" content="MhVJ4YhH6BFMk3VMQ_okf6dDKl64kAu05XQt9F__0qY"/>
<meta name="robots" content="index,follow"/>
<meta name="format-detection" content="telephone=no"/>
<meta name="generator" content="FirstSpirit 5.2.240809.5a859e4"/>
<link href="/media/resources/css_2/app_1.1.8.css" rel="stylesheet"/>
<link href="/media/resources/css_2/themes_css/theme-9c.css" rel="stylesheet"/>
<link rel="shortcut icon" type="image/x-icon" href="/media/resources/images_1/favicon_48x48.png">
<link rel="icon" type="image/png" href="/media/resources/images_1/favicon_48x48.png" sizes="48x48">
<link rel="icon" type="image/png" href="/media/resources/images_1/favicon_80x80.png" sizes="80x80">
<link rel="apple-touch-icon" sizes="80x80" href="/media/resources/images_1/favicon_80x80.png">
</head>
<body>
```

#1.3 HTML tags

```
<html>
  <head>
  </head>
  <body>
    <p>Here’s a paragraph of text!</p>
    <p>Here's a second paragraph of text!</p>
  </body>
</html>
```

Text inside <p> tags is displayed as separate paragraphs

#1.3 HTML tags



Here's a paragraph of text!

Here's a second paragraph of text!

#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>
```

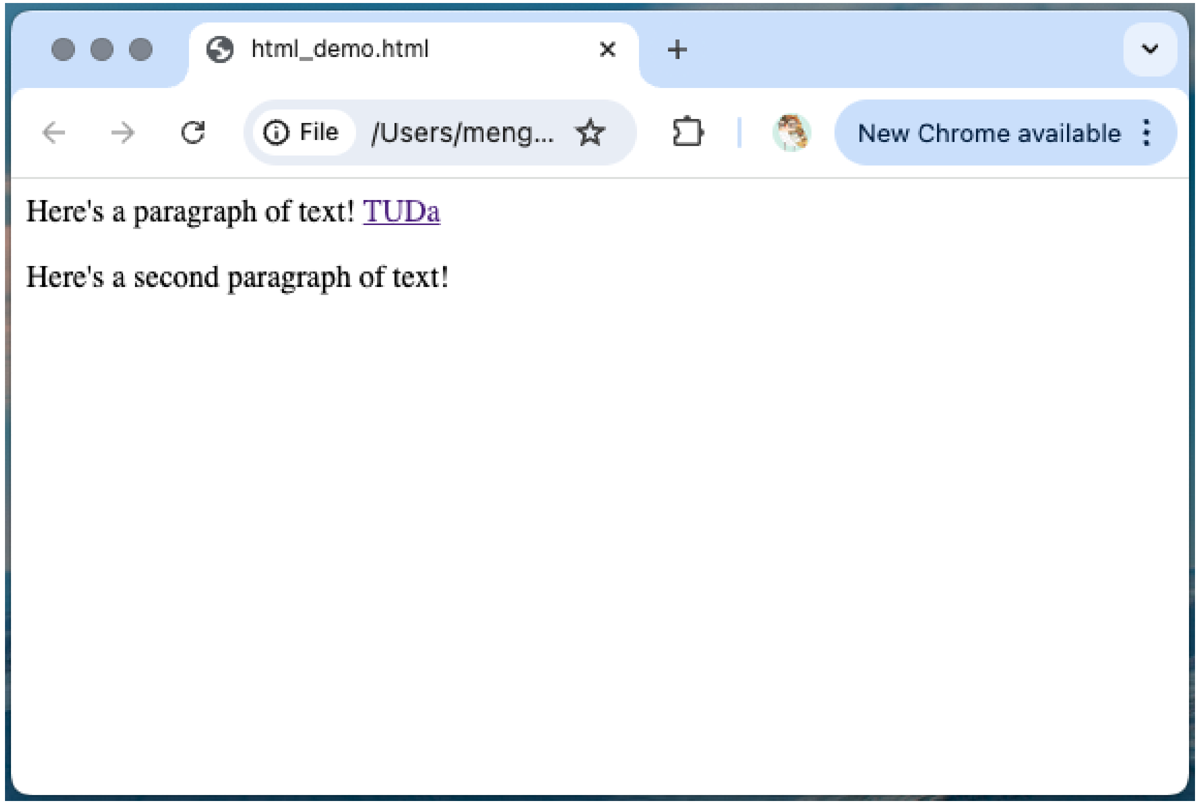
#1.3 HTML tags

```
<p>

    Here's a paragraph of text!

    <a href="https://www.tu-darmstadt.de">TUDa</a>

</p>
```

→ defines a hyperlink

href attribute specifies the destination URL

#1.3 HTML tags



Here's a paragraph of text! TUDa

Here's a second paragraph of text!

#1.3 HTML tags

Common HTML tags

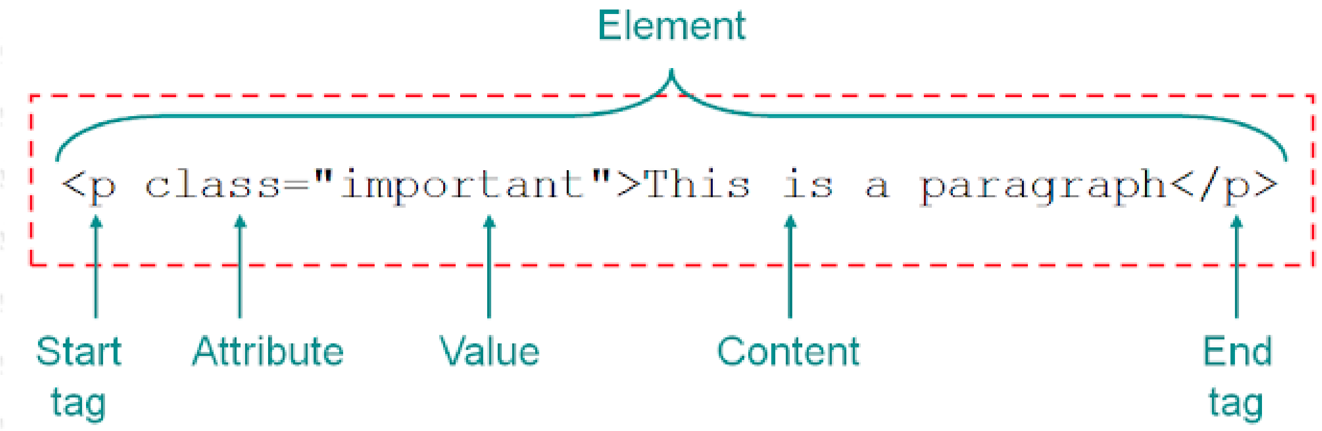
- <div>: Defines a division or a section
- <b>: Makes text bold
- <i>: Italicizes text
- <table>: Creates a table structure
- <form>: Defines an input form

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. <a href="URL">



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

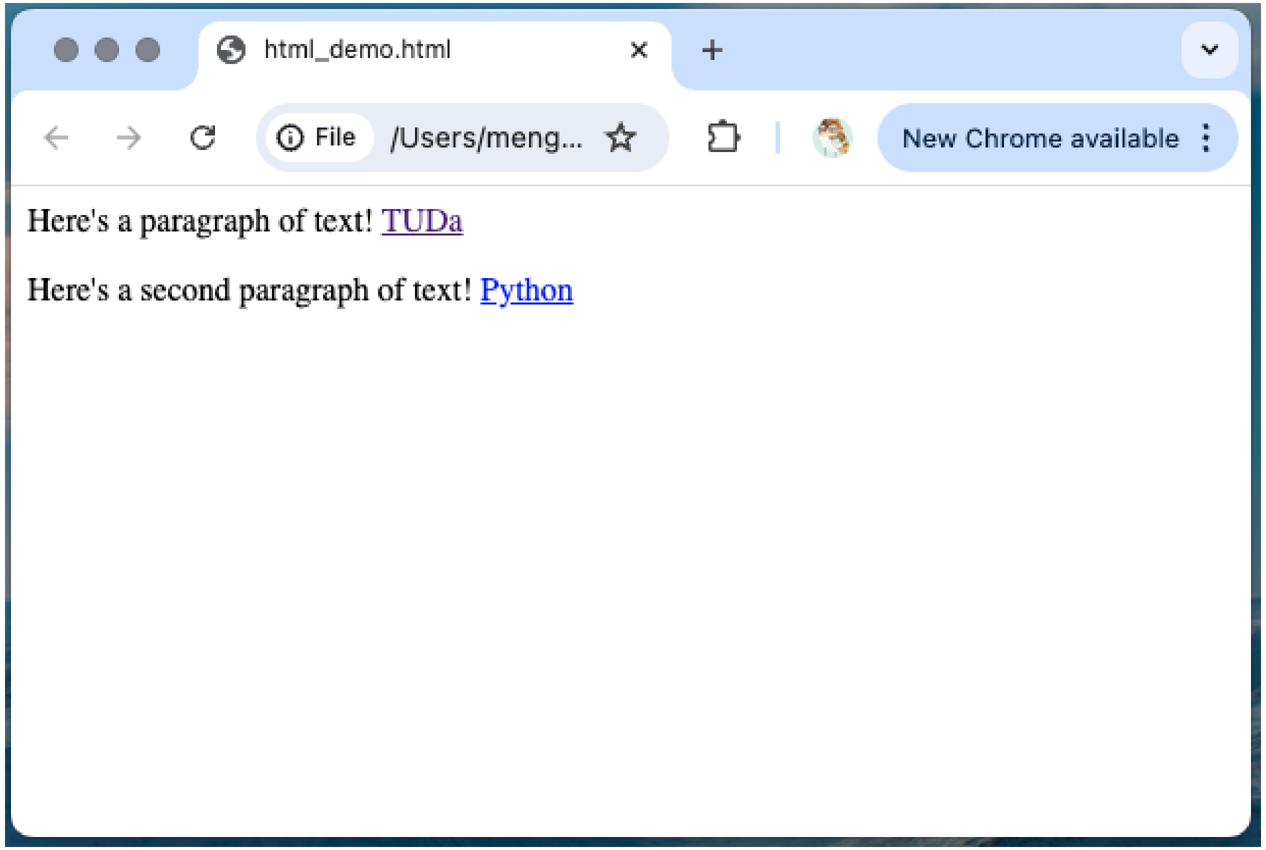
#1.4 HTML attributes

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



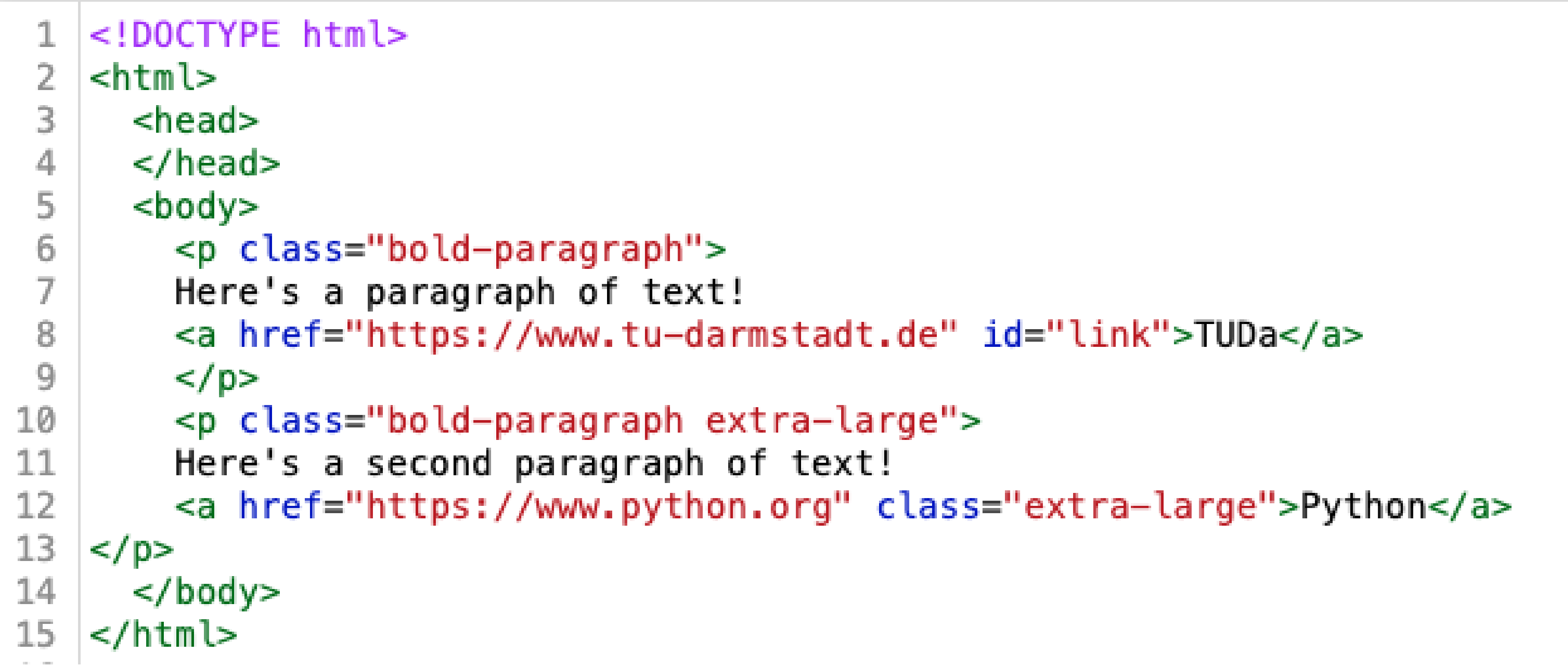
Here's a paragraph of text! TUDa

Here's a second paragraph of text! Python

#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



```
<!DOCTYPE html>
<html>
  <head>
  </head>
  <body>
    <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>
  </body>
</html>
```

#1.6 Summary of web page basics

- 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

#1 Web page basics

#2 Web scraping with Python

#2.1 What is web scraping?

- 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?

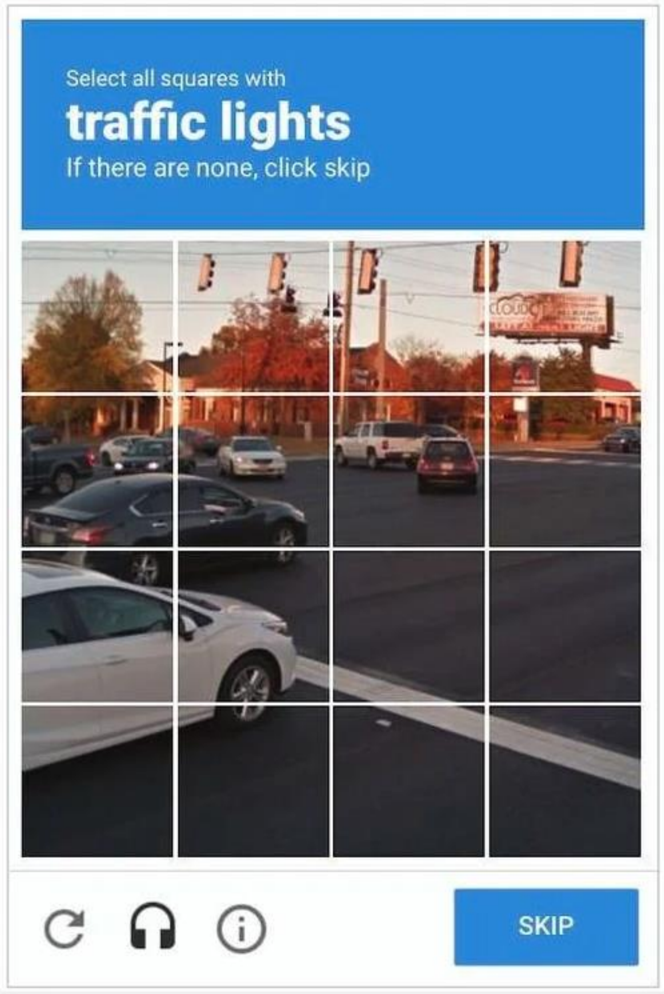
Web scraping is particularly useful in scenarios where data is needed but not readily available in structured formats.

- 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

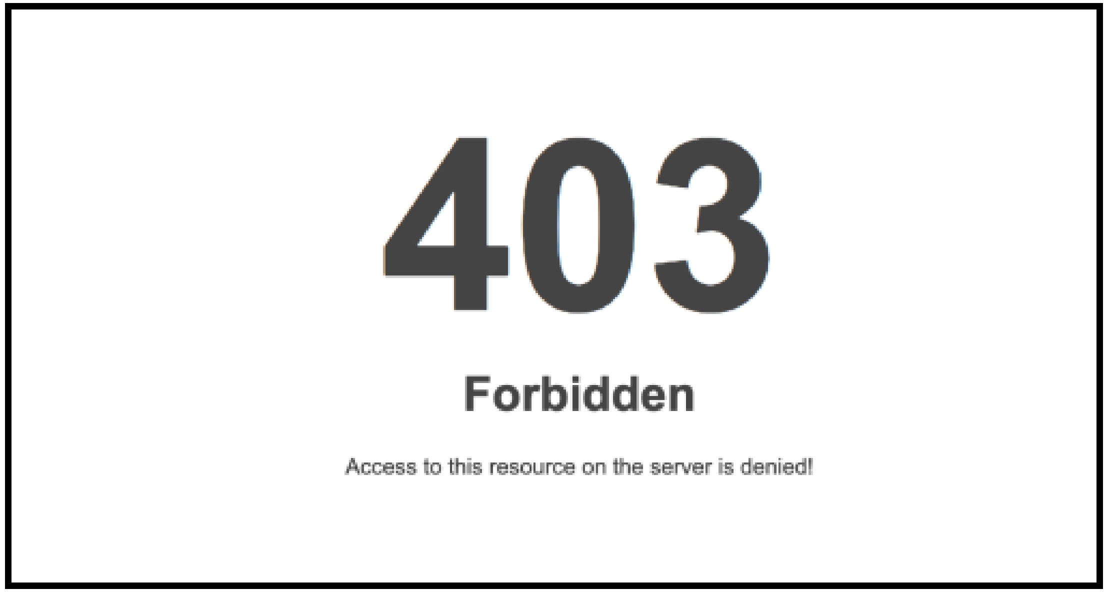
- Websites using JavaScript to load content dynamically
- IP Blocking and CAPTCHAs
- Changing Website Structures



Select all squares with
traffic lights
If there are none, click skip

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

#2.4 Legal and ethical considerations



403

Forbidden

Access to this resource on the server is denied!

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

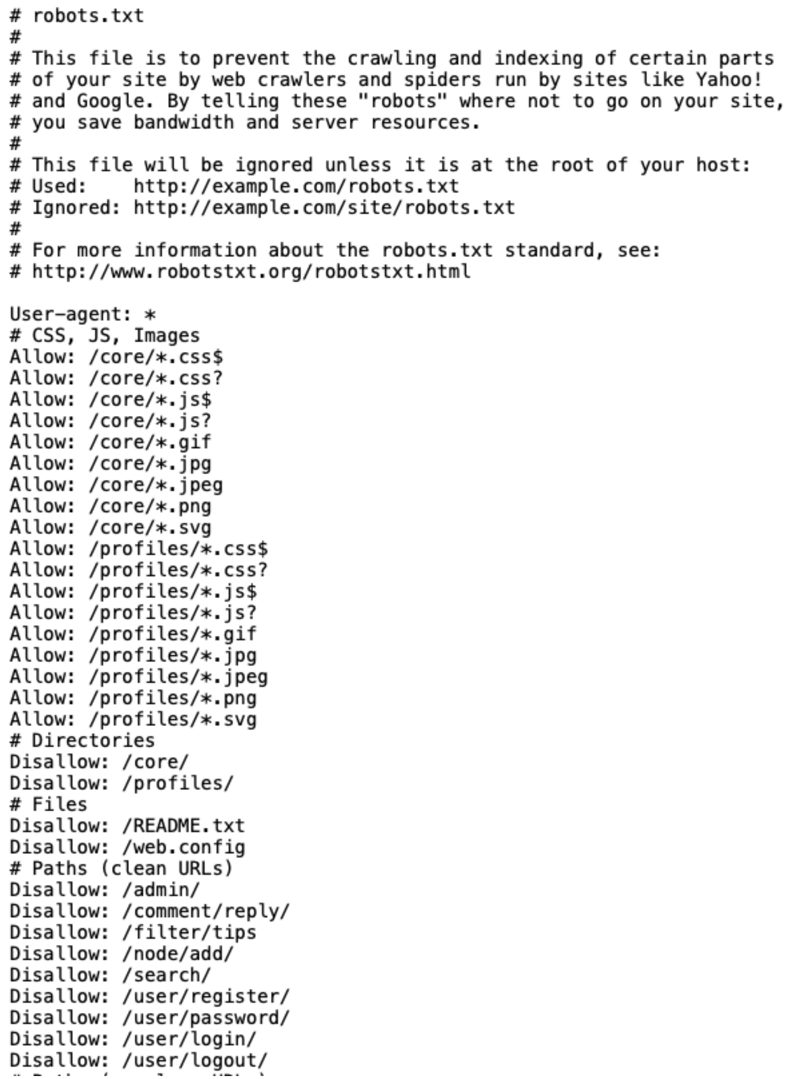
Legal

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

Ethical

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

#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/
```

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

2. Read the second paragraph in “About /robots.txt”

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

#2.6 Activity: let’s scrape a page

Please go to the Jupyter notebook.

#2.7 Summary of web scraping with Python

- 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

#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