HTTP Requests

1405 Instructor: Ruiqing (Sam) Cao

Required Python Libraries for Today

Main libraries: json, requests, csv, Beautiful Soup (bs4), pyjsonviewer

```
import requests
from bs4 import BeautifulSoup
import pyjsonviewer
import json
import csv
```

Recurring libraries (we’ll see a lot more of later): numpy, pandas, matplotlib

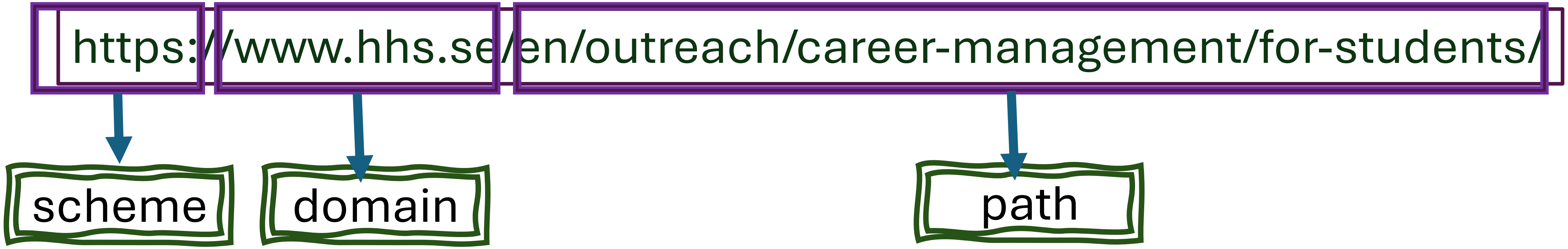
```
import numpy as np
import pandas as pd
from matplotlib import pyplot as plt
```

If a `module` is not yet installed, you can type `%pip install module` inline in your Jupyter Notebook to install it

Uniform Resource Locator (URL)

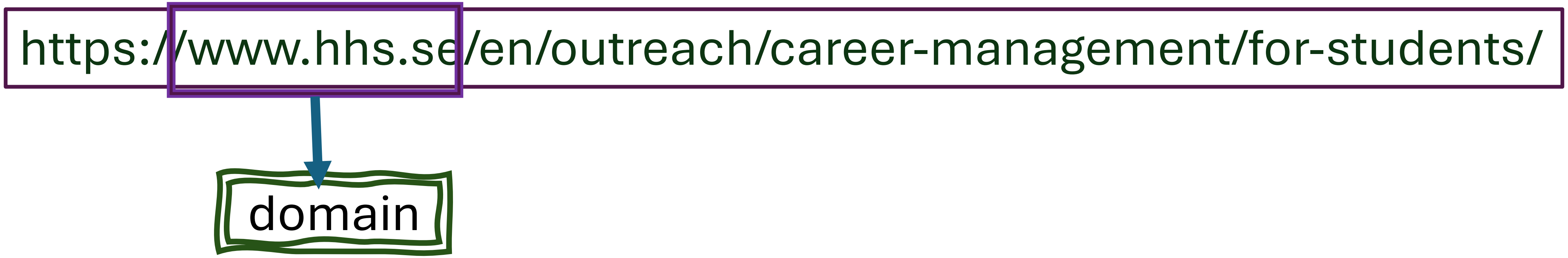
- URLs are a standardized way to identify an Internet-based resource (e.g., a website)

- Webpage addresses are expressed as URLs:



https://www.hhs.se/en/outreach/career-management/for-students/

Uniform Resource Locator (URL)



https://www.hhs.se/en/outreach/career-management/for-students/

domain

- Top-level domain: “se”
- Second-level domain: “hhs”
- Third-level domain: “www”

Can have multiple third-level domains for the same company’s website, e.g., “community.spotify.com”, “developer.spotify.com”, etc.

HTTP Request: Response Status Code

```
import requests
response = requests.get('https://www.hhs.se/en/outreach/career-
management/for-students'
response.status_code
```

This is going to return 200
200: OK

If the input URL is invalid, you'll get the client error of 404
404: Not Found

Status Code	
1xx	Informational
2xx	**Successful**
3xx	Redirect
4xx	**Client error**
5xx	Server error

HTTP Request: Response Headers

```
import requests
response = requests.get('https://www.hhs.se/en/outreach/career-management/for-students'
response.headers
```

```
{'Cache-Control': 'private', 'Content-Type':
'text/html; charset=utf-8', 'Content-Encoding':
'gzip', 'Vary': 'Accept-Encoding', 'Server':
'Microsoft-IIS/10.0', 'Set-Cookie':
'ASP.NET_SessionId=ytuxhmjqz00u504vw4l44ptq; path=/;
HttpOnly; SameSite=Lax', 'X-AspNetMvc-Version': '5.2',
'X-AspNet-Version': '4.0.30319', 'X-Content-Type-
Options': 'nosniff', 'Strict-Transport-Security':
'max-age=63072000; includeSubDomains', 'Feature-
Policy': 'fullscreen *', 'Date': 'Fri, 06 Dec 2024
13:41:40 GMT', 'Content-Length': '37608'}
```

Date the request was sent:
`'Fri, 06 Dec 2024 13:41:40 GMT'`

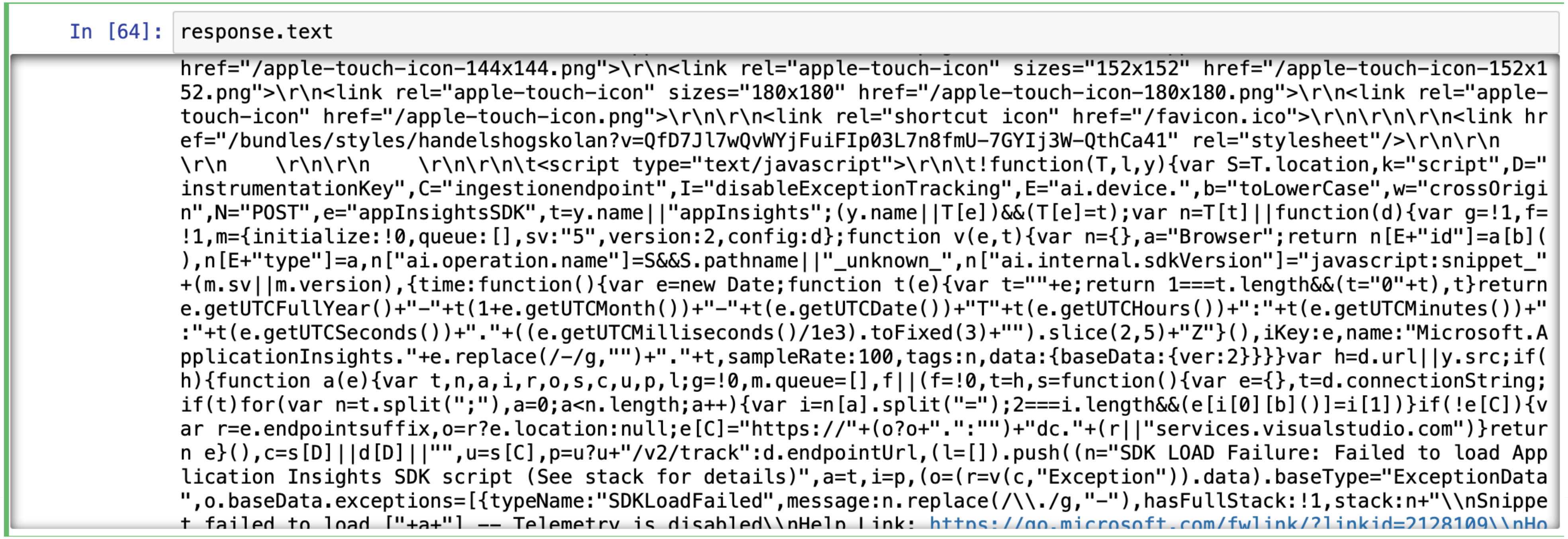
Type of data in response:
`'text/html; charset=utf-8'`

➔ Thus, we can use the attribute `response.text` to display the body of the HTML webpage

HTTP Request: Response Text

```
import requests
response = requests.get('https://www.hhs.se/en/outreach/career-management/for-students'
```

Get the body of the HTML page:



```
In [64]: response.text
```

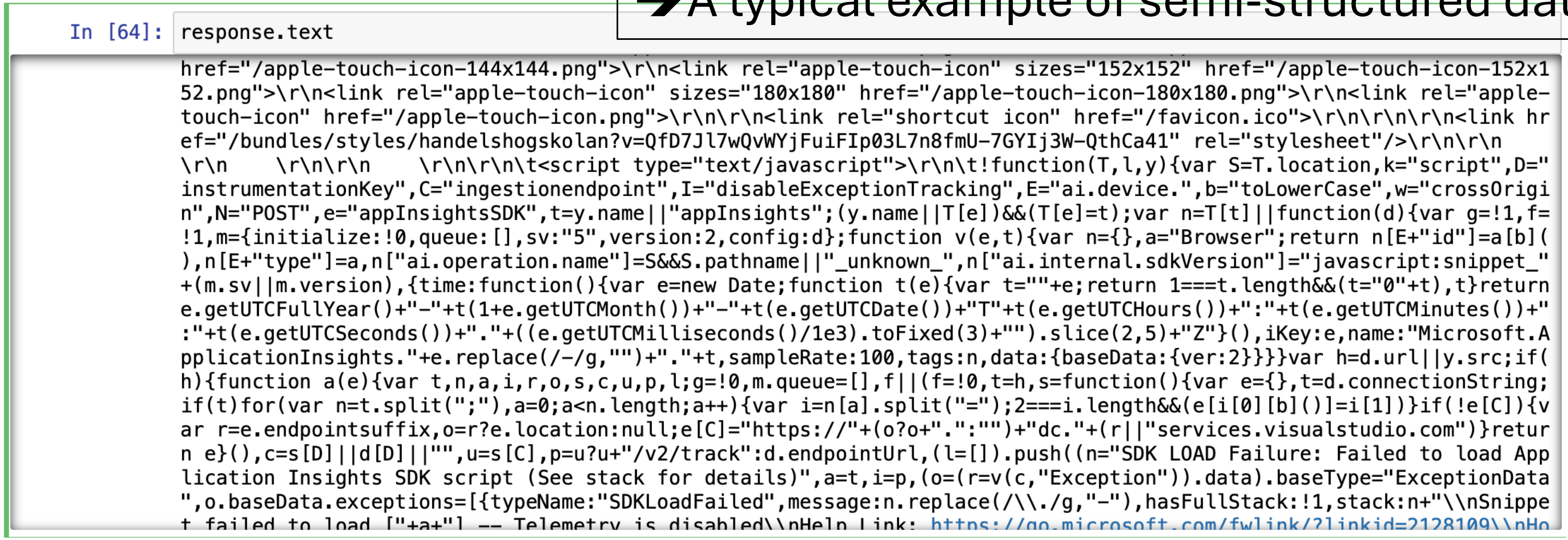
```
href="/apple-touch-icon-144x144.png">\r\n<link rel="apple-touch-icon" sizes="152x152" href="/apple-touch-icon-152x1
52.png">\r\n<link rel="apple-touch-icon" sizes="180x180" href="/apple-touch-icon-180x180.png">\r\n<link rel="apple-
touch-icon" href="/apple-touch-icon.png">\r\n\r\n<link rel="shortcut icon" href="/favicon.ico">\r\n\r\n\r\n<link hr
ef="/bundles/styles/handelshogskolan?v=QfD7Jl7wQvWYjFuiFIp03L7n8fmU-7GYIj3W-QthCa41" rel="stylesheet"/>\r\n\r\n
\r\n    \r\n\r\n    \r\n\r\n\t<script type="text/javascript">\r\n\t!function(T,l,y){var S=T.location,k="script",D="
instrumentationKey",C="ingestionendpoint",I="disableExceptionTracking",E="ai.device.",b="toLowerCase",w="crossOrigi
n",N="POST",e="appInsightsSDK",t=y.name||"appInsights";(y.name||T[e])&&(T[e]=t);var n=T[t]||function(d){var g=!1,f=
!1,m={initialize:!0,queue:[],sv:"5",version:2,config:d};function v(e,t){var n={},a="Browser";return n[E+"id"]=a[b](
),n[E+"type"]=a,n["ai.operation.name"]=S&&S.pathname||"_unknown_",n["ai.internal.sdkVersion"]="javascript:snippet_"
+(m.sv||m.version),{time:function(){var e=new Date;function t(e){var t=""+e;return 1===t.length&&(t="0"+t),t}return
e.getUTCFullYear()+"-"+t(1+e.getUTCMonth())+"-"+t(e.getUTCDate())+"T"+t(e.getUTCHours())+":"+t(e.getUTCMinutes())+"
:"+t(e.getUTCSeconds())+"."+((e.getUTCMilliseconds()/1e3).toFixed(3)+"").slice(2,5)+"Z"}(),iKey:e,name:"Microsoft.A
pplicationInsights."+e.replace(/-/g,"")+"."+t,sampleRate:100,tags:n,data:{baseData:{ver:2}}}}var h=d.url||y.src;if(
h){function a(e){var t,n,a,i,r,o,s,c,u,p,l;g=!0,m.queue=[],f||(f=!0,t=h,s=function(){var e={},t=d.connectionString;
if(t)for(var n=t.split(";"),a=0;a<n.length;a++){var i=n[a].split("=");2===i.length&&(e[i[0][b]()]=i[1])}if(!e[C]){v
ar r=e.endpointsuffix,o=r?e.location:null;e[C]="https://"+(o?o+".":"")+"dc."+(r||"services.visualstudio.com")}retur
n e}(),c=s[D]||d[D]||"",u=s[C],p=u?u+"/v2/track":d.endpointUrl,(l=[]).push((n="SDK LOAD Failure: Failed to load App
lication Insights SDK script (See stack for details)",a=t,i=p,(o=(r=v(c,"Exception")).data).baseType="ExceptionData
",o.baseData.exceptions=[{typeName:"SDKLoadFailed",message:n.replace(/\\./g,"-"),hasFullStack:!1,stack:n+"\\nSnippe
t failed to load ["+a+"] -- Telemetry is disabled\\nHelp Link: https://go.microsoft.com/fwlink/?linkid=2128109\\nHo
```

HTTP Request: Response Text (HTML)

```
import requests
response = requests.get('https://www.hhs.se/en/outreach/career-management/for-students'
```

Get the body of the HTML page:

➔A typical example of semi-structured data



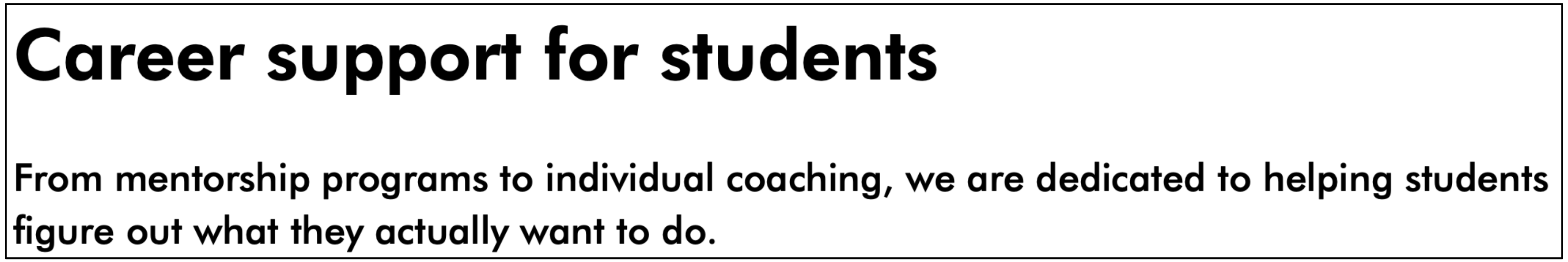
```
In [64]: response.text
```

```
href="/apple-touch-icon-144x144.png">\r\n<link rel="apple-touch-icon" sizes="152x152" href="/apple-touch-icon-152x1
52.png">\r\n<link rel="apple-touch-icon" sizes="180x180" href="/apple-touch-icon-180x180.png">\r\n<link rel="apple-
touch-icon" href="/apple-touch-icon.png">\r\n\r\n<link rel="shortcut icon" href="/favicon.ico">\r\n\r\n\r\n<link hr
ef="/bundles/styles/handelshogskolan?v=QfD7Jl7wQvWYjFuiFIp03L7n8fmU-7GYIj3W-QthCa41" rel="stylesheet"/>\r\n\r\n
\r\n    \r\n\r\n    \r\n\r\n\t<script type="text/javascript">\r\n\t!function(T,l,y){var S=T.location,k="script",D="
instrumentationKey",C="ingestionendpoint",I="disableExceptionTracking",E="ai.device.",b="toLowerCase",w="crossOrigi
n",N="POST",e="appInsightsSDK",t=y.name||"appInsights";(y.name||T[e])&&(T[e]=t);var n=T[t]||function(d){var g=!1,f=
!1,m={initialize:!0,queue:[],sv:"5",version:2,config:d};function v(e,t){var n={},a="Browser";return n[E+"id"]=a[b](
),n[E+"type"]=a,n["ai.operation.name"]=S&&S.pathname||"_unknown_",n["ai.internal.sdkVersion"]="javascript:snippet_"
+(m.sv||m.version),{time:function(){var e=new Date;function t(e){var t=""+e;return 1===t.length&&(t="0"+t),t}return
e.getUTCFullYear()+"-"+t(1+e.getUTCMonth())+"-"+t(e.getUTCDate())+"T"+t(e.getUTCHours())+":"+t(e.getUTCMinutes())+"
:"+t(e.getUTCSeconds())+"."+((e.getUTCMilliseconds()/1e3).toFixed(3)+"").slice(2,5)+"Z"}(),iKey:e,name:"Microsoft.A
pplicationInsights."+e.replace(/-/g,"")+"."+t,sampleRate:100,tags:n,data:{baseData:{ver:2}}}}var h=d.url||y.src;if(
h){function a(e){var t,n,a,i,r,o,s,c,u,p,l;g=!0,m.queue=[],f||(f=!0,t=h,s=function(){var e={},t=d.connectionString;
if(t)for(var n=t.split(";"),a=0;a<n.length;a++){var i=n[a].split("=");2===i.length&&(e[i[0][b]()]=i[1])}if(!e[C]){v
ar r=e.endpointsuffix,o=r?e.location:null;e[C]="https://"+(o?o+".":"")+"dc."+(r||"services.visualstudio.com")}retur
n e}(),c=s[D]||d[D]||"",u=s[C],p=u?u+"/v2/track":d.endpointUrl,(l=[]).push((n="SDK LOAD Failure: Failed to load App
lication Insights SDK script (See stack for details)",a=t,i=p,(o=(r=v(c,"Exception")).data).baseType="ExceptionData
",o.baseData.exceptions=[{typeName:"SDKLoadFailed",message:n.replace(/\\./g,"-"),hasFullStack:!1,stack:n+"\\nSnippe
t failed to load ["+a+"] -- Telemetry is disabled\\nHelp Link: https://go.microsoft.com/fwlink/?linkid=2128109\\nHo
```

Let's explore its structure further...

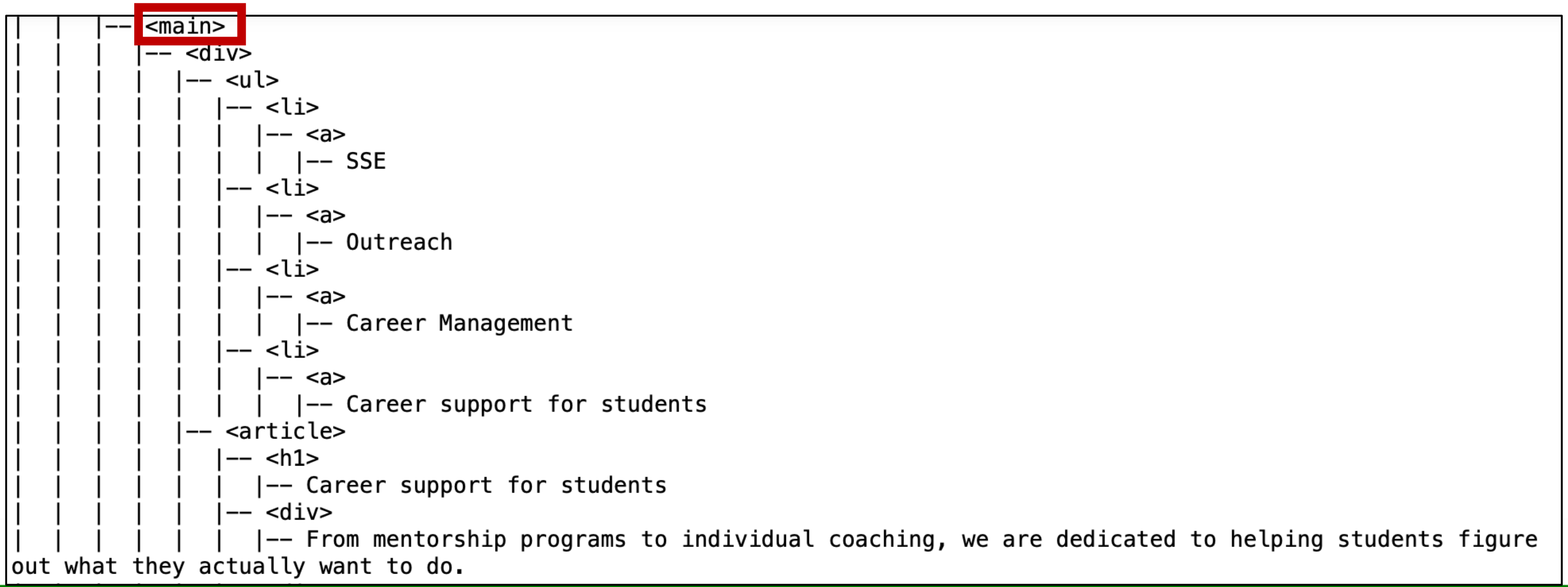
HTTP Request: HTML Tree Structure

Print the tree structure, and the first 500 character of every piece of text within the HTML page:



Career support for students

From mentorship programs to individual coaching, we are dedicated to helping students figure out what they actually want to do.

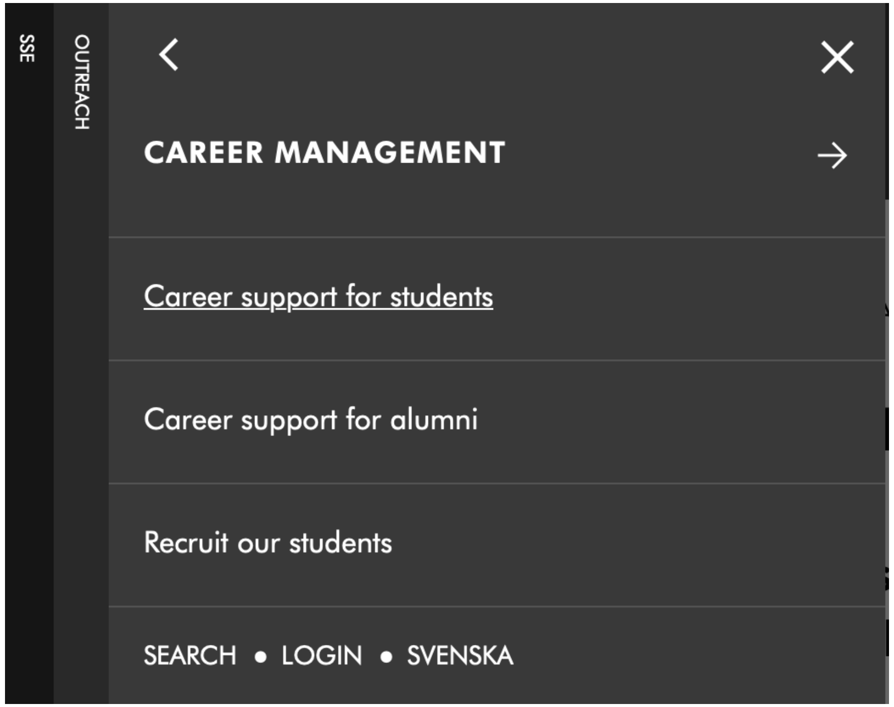


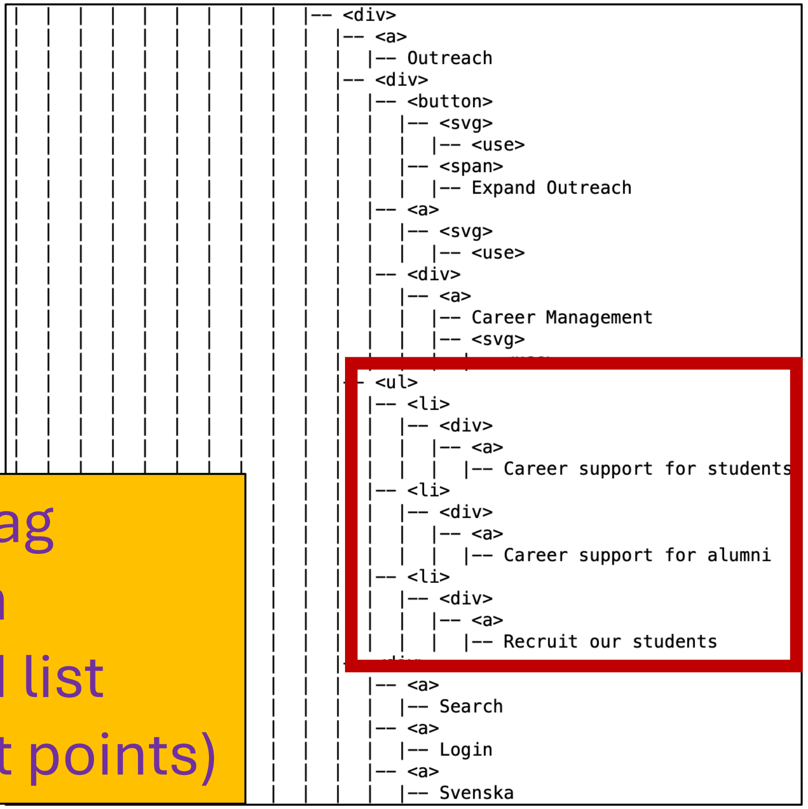
```
|   |   |-- <main>
|   |   |   |-- <div>
|   |   |   |   |-- <ul>
|   |   |   |   |   |-- <li>
|   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |-- SSE
|   |   |   |   |   |-- <li>
|   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |-- Outreach
|   |   |   |   |   |-- <li>
|   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |-- Career Management
|   |   |   |   |   |-- <li>
|   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |-- Career support for students
|   |   |   |   |-- <article>
|   |   |   |   |   |-- <h1>
|   |   |   |   |   |   |-- Career support for students
|   |   |   |   |   |-- <div>
|   |   |   |   |   |   |-- From mentorship programs to individual coaching, we are dedicated to helping students figure
out what they actually want to do.
```

The <main> tag specifies the main content of the HTML document and should be unique.

HTTP Request: HTML Tree Structure

Print the tree structure, and the first 500 character of every piece of text within the HTML page:



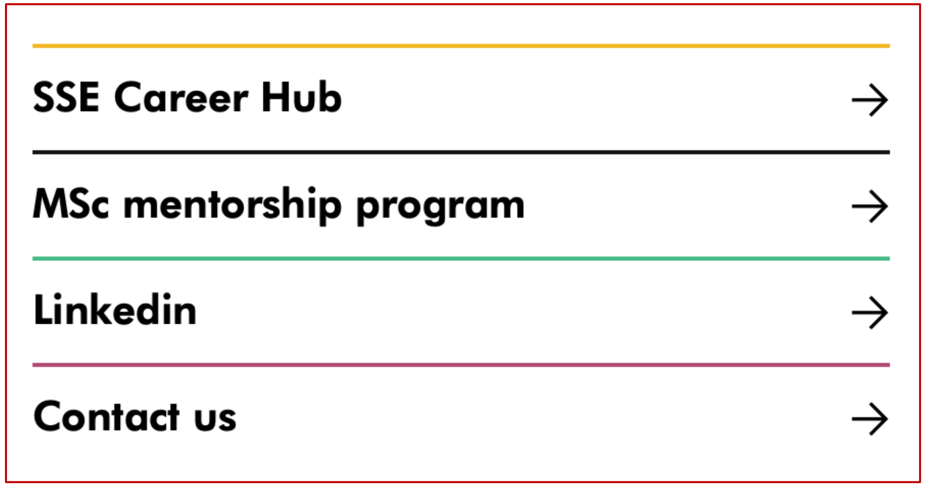


```
|   |   |   |   |   |   |   |   |   |-- <div>
|   |   |   |   |   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |   |   |   |   |-- Outreach
|   |   |   |   |   |   |   |   |   |   |-- <div>
|   |   |   |   |   |   |   |   |   |   |   |-- <button>
|   |   |   |   |   |   |   |   |   |   |   |   |-- <svg>
|   |   |   |   |   |   |   |   |   |   |   |   |   |-- <use>
|   |   |   |   |   |   |   |   |   |   |   |   |-- <span>
|   |   |   |   |   |   |   |   |   |   |   |   |   |-- Expand Outreach
|   |   |   |   |   |   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |   |   |   |   |   |-- <svg>
|   |   |   |   |   |   |   |   |   |   |   |   |   |-- <use>
|   |   |   |   |   |   |   |   |   |   |   |-- <div>
|   |   |   |   |   |   |   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |   |   |   |   |   |   |-- Career Management
|   |   |   |   |   |   |   |   |   |   |   |   |   |-- <svg>
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |-- <use>
|   |   |   |   |   |   |   |   |   |   |-- <ul>
|   |   |   |   |   |   |   |   |   |   |   |-- <li>
|   |   |   |   |   |   |   |   |   |   |   |   |-- <div>
|   |   |   |   |   |   |   |   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |-- Career support for students
|   |   |   |   |   |   |   |   |   |   |   |-- <li>
|   |   |   |   |   |   |   |   |   |   |   |   |-- <div>
|   |   |   |   |   |   |   |   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |-- Career support for alumni
|   |   |   |   |   |   |   |   |   |   |   |-- <li>
|   |   |   |   |   |   |   |   |   |   |   |   |-- <div>
|   |   |   |   |   |   |   |   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |   |   |   |   |   |   |   |-- Recruit our students
|   |   |   |   |   |   |   |   |   |   |-- <div>
|   |   |   |   |   |   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |   |   |   |   |   |-- Search
|   |   |   |   |   |   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |   |   |   |   |   |-- Login
|   |   |   |   |   |   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |   |   |   |   |   |-- Svenska
```

The <ul> tag defines an unordered list (i.e., bullet points)

HTTP Request: HTML Tree Structure

Print the tree structure, and the first 500 character of every piece of text within the HTML page:

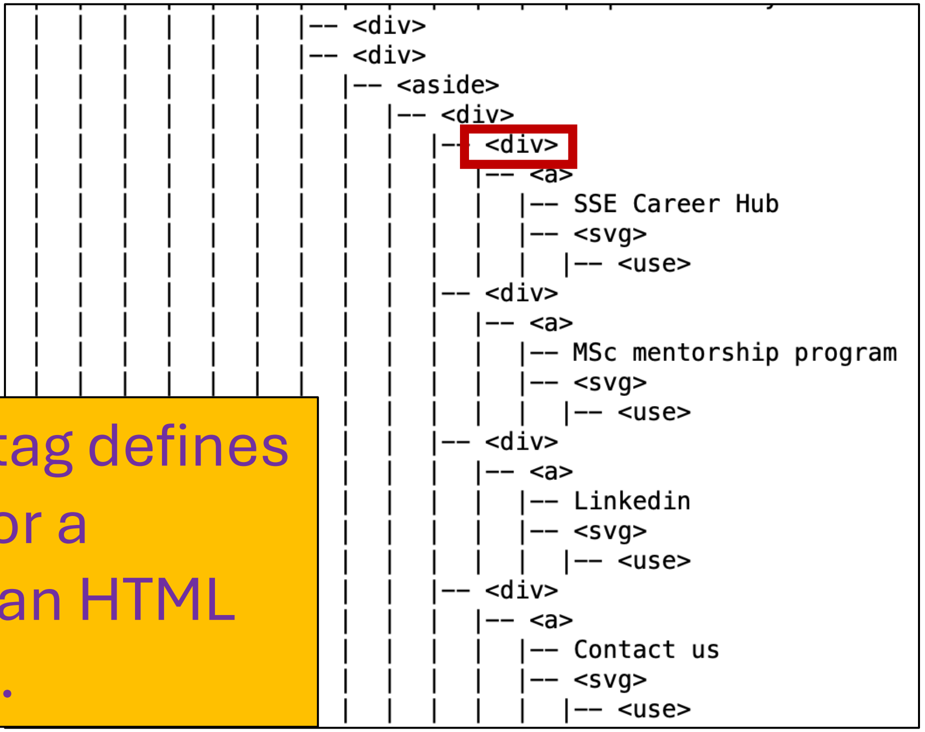


SSE Career Hub →

MSc mentorship program →

Linkedin →

Contact us →



```
|   |   |   |   |   |   |-- <div>
|   |   |   |   |   |   |-- <div>
|   |   |   |   |   |   |   |-- <aside>
|   |   |   |   |   |   |   |   |-- <div>
|   |   |   |   |   |   |   |   |   |-- <div>
|   |   |   |   |   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |   |   |   |   |-- SSE Career Hub
|   |   |   |   |   |   |   |   |   |   |   |-- <svg>
|   |   |   |   |   |   |   |   |   |   |   |   |-- <use>
|   |   |   |   |   |   |   |   |   |-- <div>
|   |   |   |   |   |   |   |   |   |   |-- <a>
|   |   |   |   |   |   |   |   |   |   |   |-- MSc mentorship program
|   |   |   |   |   |   |   |   |   |   |   |-- <svg>
                            |   |   |   |   |   |-- <use>
                            |   |   |-- <div>
                            |   |   |   |-- <a>
                            |   |   |   |   |-- Linkedin
                            |   |   |   |   |-- <svg>
                            |   |   |   |   |   |-- <use>
                            |   |   |-- <div>
                            |   |   |   |-- <a>
                            |   |   |   |   |-- Contact us
                            |   |   |   |   |-- <svg>
                            |   |   |   |   |   |-- <use>
```

The <div> tag defines a division or a section in an HTML document.