Introduction to Web Scraping for Academic Research

Jihed Ncib
PhD Candidate

School of Politics and International Relations
University College Dublin

Overview

- Process automation to extract data from websites
  - Introductory slides, definitions, and guides
  - Exercises and practice on demonstration websites

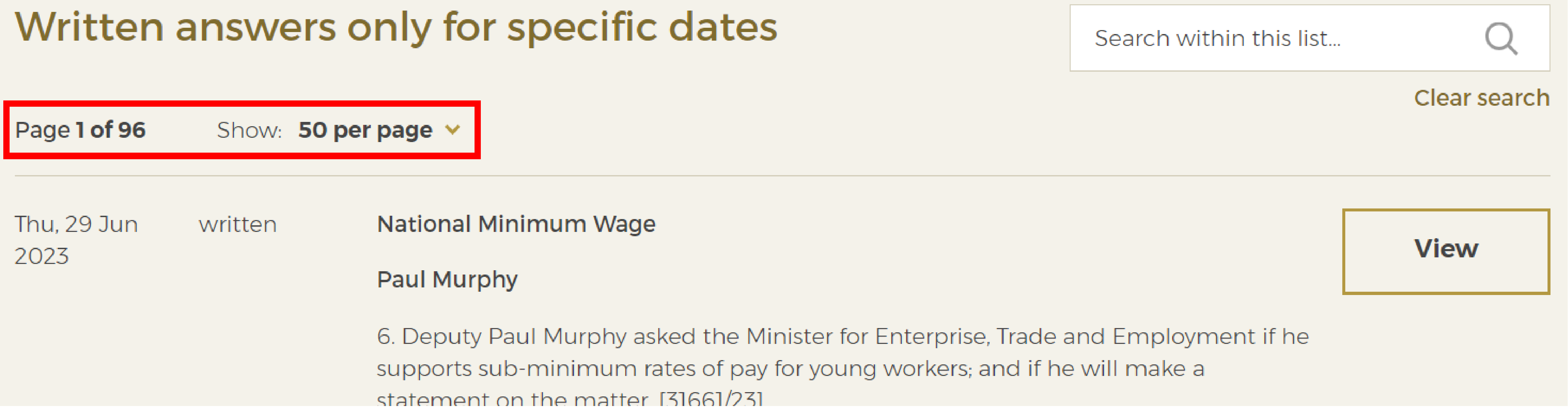
- Designed for researchers with basic knowledge of R
  - Quick reminders and brushing up on the required packages and functions

- Examples from actual research projects and real-world applications

- Follow in real time here: **github.com/jihedncib/CCSS_Workshop**

Motivation and Aims

- My own PhD dissertation examines the political communication of members of Parliament in different countries (tweets, parliamentary questions, speeches, etc.).
- Massive amounts of data would be needed and would take a lot of time if collected manually. Example: Parliamentary questions from the website of the Irish parliament: Around 4800 questions in one month (June 2023).



Written answers only for specific dates

Page 1 of 96 Show: 50 per page

Thu, 29 Jun 2023 written

National Minimum Wage

Paul Murphy

6. Deputy Paul Murphy asked the Minister for Enterprise, Trade and Employment if he supports sub-minimum rates of pay for young workers; and if he will make a statement on the matter. [31661/23]

Installing the Required Tools

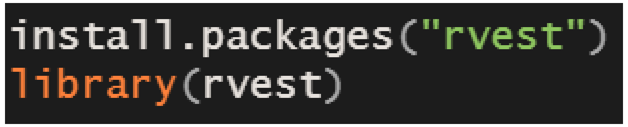
- R – Download and install
  - created for data analysis, extending for other purposes e.g., accessing websites
  - allows for all three steps in one environment: accessing websites, scraping data, and processing data

- Download R from https://cloud.r-project.org

- Download RStudio from https://rstudio.com/products/rstudio/download
  - integrated development environment (IDE) for R

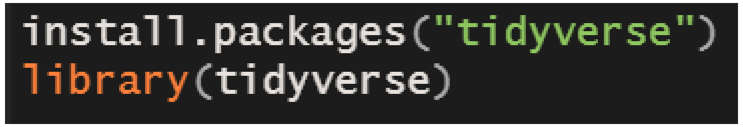
Required Packages

- ‘rvest’: Wrappers around the 'xml2' and 'httr' packages to make it easy to download, then manipulate, HTML and XML.



```
install.packages("rvest")
library(rvest)
```

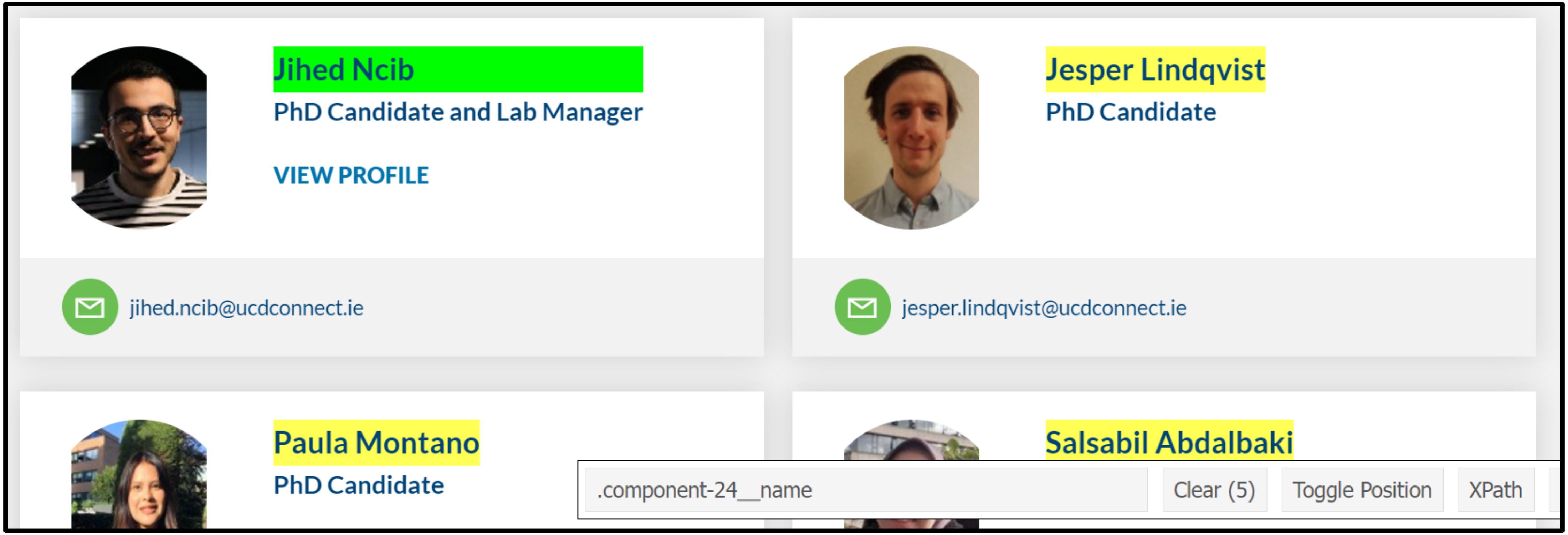
- ‘tidyverse’: A compilation of packages used to manipulate and wrangle different formats of data. We’ll mostly be using ‘dplyr’ in this workshop to clean web scraped data (that usually comes in a messy format).



```
install.packages("tidyverse")
library(tidyverse)
```

SelectorGadget

- An extension for Chrome
  - facilitates selecting what to scrape from a webpage
  - optional, but highly recommended
- Add the extension to your browser
  - search for it on Chrome's webstore https://chrome.google.com/webstore/category/extensions



Jihed Ncib
PhD Candidate and Lab Manager
jihed.ncib@ucdconnect.ie

Jesper Lindqvist
PhD Candidate
jesper.lindqvist@ucdconnect.ie

Paula Montano
PhD Candidate

Salsabil Abdalbaki

.component-24__name

Ethical Considerations

- Web scraping might be **illegal**
  - depending on who is scraping what, why, how — and under which jurisdiction
  - reflect, and check, before you scrape

- Web scraping might be more likely to be illegal if, for example,
  - it is harmful to the source commercially and/or physically
    - e.g., scraping a commercial website to create a rival website
    - e.g., scraping a website so hard and fast that it collapses

  - it gathers data that is
    - under copyright
    - not meant for the public to see
    - then used for financial gain

Ethical Considerations

- Web scraping might be **unethical**
  - depending on who is scraping what, why, and how
  - reflect before you scrape

- Web scraping might be more likely to be unethical if, for example,
  - it is — edging towards — being illegal
  - it does not respect the restrictions as defined in *robots.txt* files

  - it harvests data
    - that is otherwise available to download, e.g., through APIs
    - without purpose, at dangerous speed, repeatedly

robots.txt

- Most websites declare a robots exclusion protocol
  - making their rules known with respect to programmatic access
    - who is (not) allowed to scrape what, and sometimes, at what speed

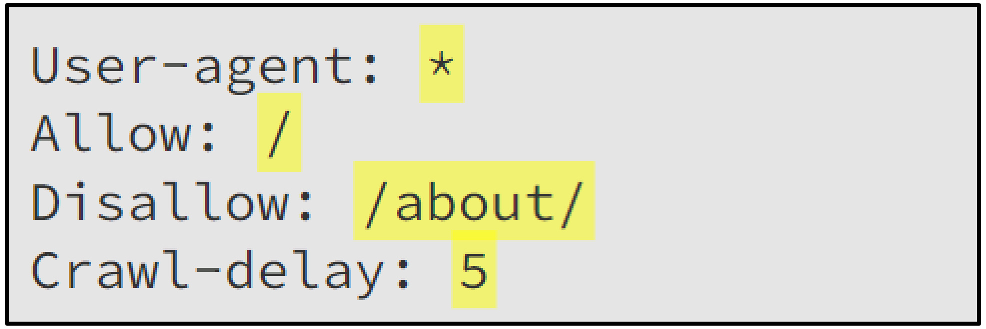
- within robots.txt files
  - available at, e.g., www.websiteurl.com/robots.txt

- The rules in robots.txt cannot not enforced upon scrapers
  - but should be respected for ethical reasons
  - https://www.washingtonpost.com/robots.txt
  - https://twitter.com/robots.txt
  - https://www.tripadvisor.com/robots.txt (with a job offer for people with web scraping skills)

robots.txt - Syntax

- It has pre-defined keys, most importantly
  - ***User-agent*** indicates who the protocol is for
  - ***Allow*** indicates which part(s) of the website can be scraped
  - ***Disallow i***ndicates which part(s) must not be scraped
  - ***Crawl-delay*** indicates how fast the website could be scraped

- Example:
  - * indicates the protocol is for everyone
  - / indicates all sections and pages
  - /about/ indicates a specific path
  - values for Crawl-delay are in seconds

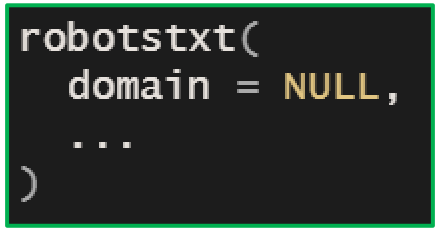


```
User-agent: *
Allow: /
Disallow: /about/
Crawl-delay: 5
```

robots.txt – robotstxt Package

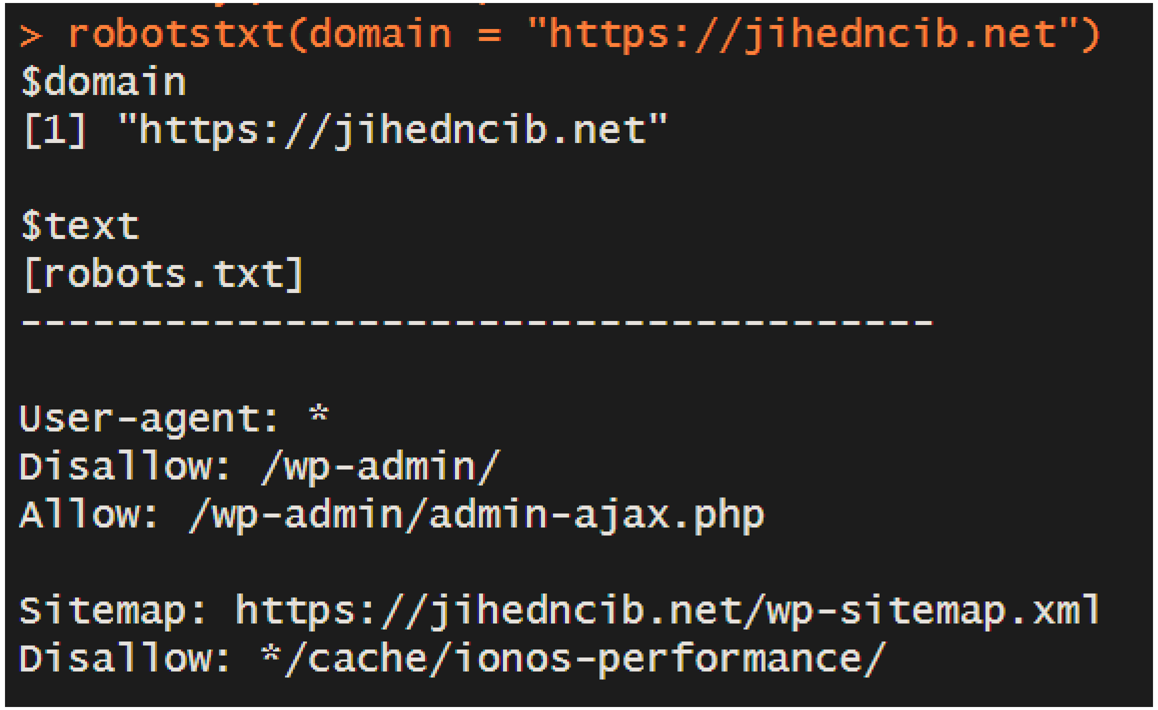
- The robotstxt packages facilitates checking website protocols
  - from within R — no need to visit websites via browser
  - provides functions to check, among others, the rules for specific paths and/or agents

- Two main functions
  - *robotstxt*, which gets complete protocols
  - *paths_allowed*, which checks protocols for one or more specific paths



```
robotstxt(
  domain = NULL,
  ...
)
```

robots.txt – robotstxt Package

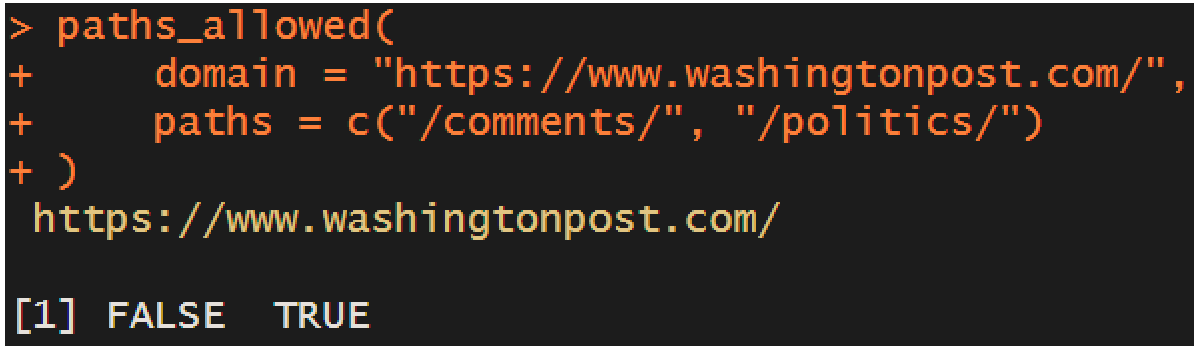


```
> robotstxt(domain = "https://jihedncib.net")
$domain
[1] "https://jihedncib.net"

$text
[robots.txt]
--------------------------------------


User-agent: *
Disallow: /wp-admin/
Allow: /wp-admin/admin-ajax.php

Sitemap: https://jihedncib.net/wp-sitemap.xml
Disallow: */cache/ionos-performance/
```



```
> paths_allowed(
+     domain = "https://www.washingtonpost.com/",
+     paths = c("/comments/", "/politics/")
+ )
 https://www.washingtonpost.com/

[1] FALSE  TRUE
```

HTML Basics – Source Code

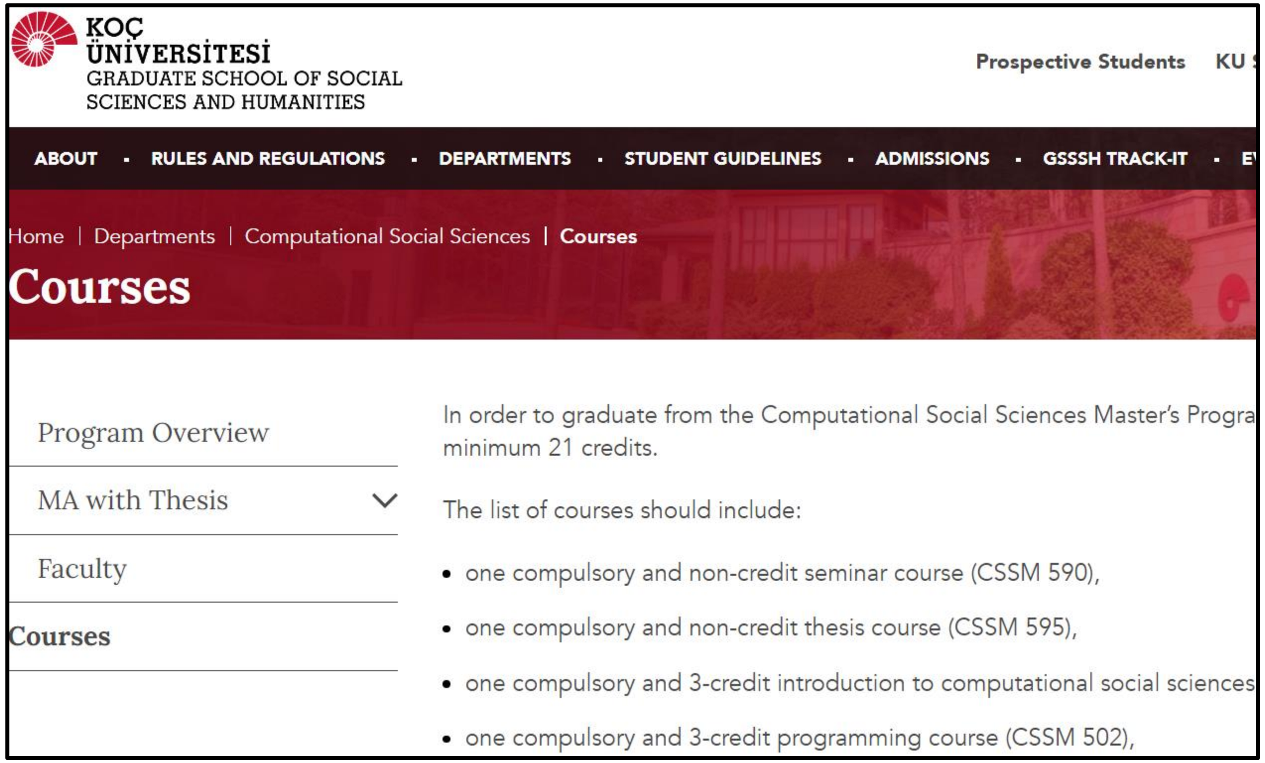
- Webpages include more than what is immediately visible to visitors
    - Code for structure, style, and functionality — interpreted by browsers first
        - HTML provides the structure
        - CSS provides the style
        - JavaScript provides functionality, if any

- Web scraping requires working with the source code
    - even when scraping only what is already visible
    - to choose one or more desired parts of the visible e.g., text in table and/or bold only

- Source code also offers more, invisible, data to be scraped
    - e.g., URLs hidden under text

HTML Basics – Source Code

- CTRL + U displays the source code of a page (Or right click > Display page source)



KOÇ
ÜNİVERSİTESİ
GRADUATE SCHOOL OF SOCIAL
SCIENCES AND HUMANITIES

Prospective Students
KU S
ABOUT · RULES AND REGULATIONS · DEPARTMENTS · STUDENT GUIDELINES · ADMISSIONS · GSSSH TRACK-IT · E
Home | Departments | Computational Social Sciences | Courses

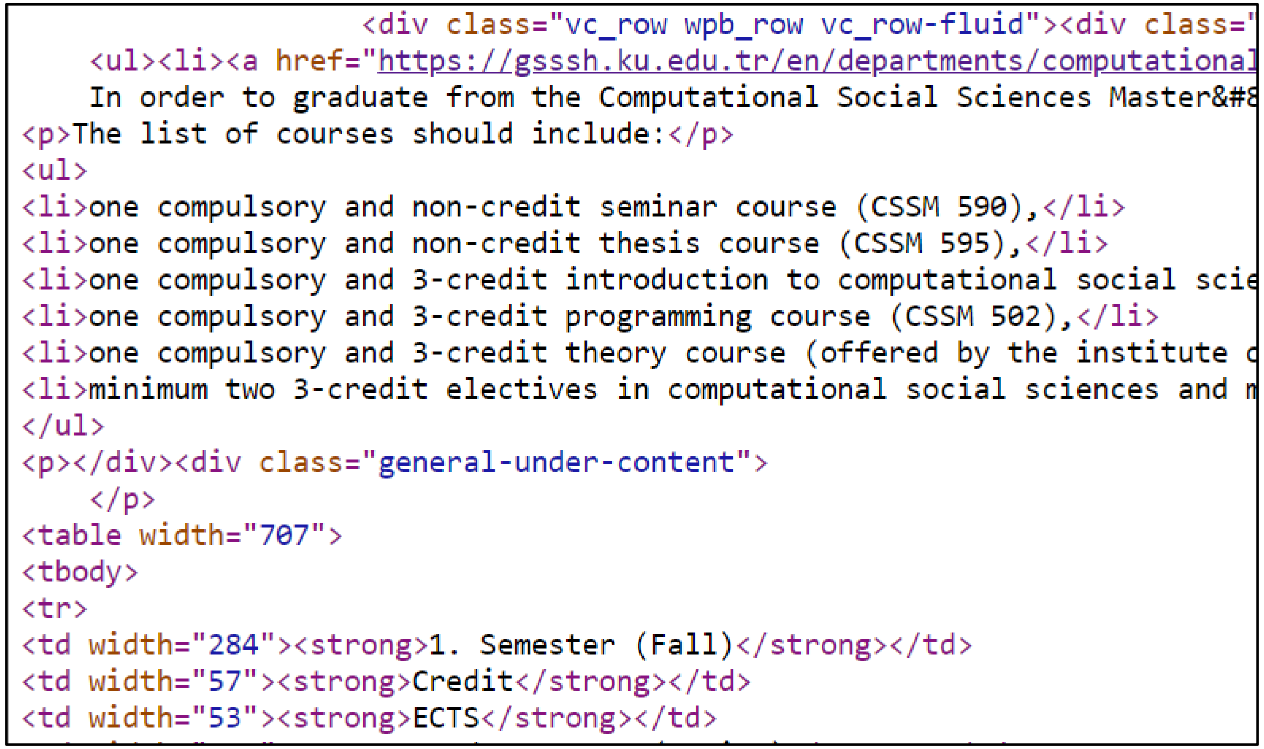
Courses

Program Overview
MA with Thesis
Faculty
Courses

In order to graduate from the Computational Social Sciences Master's Progra
minimum 21 credits.

The list of courses should include:

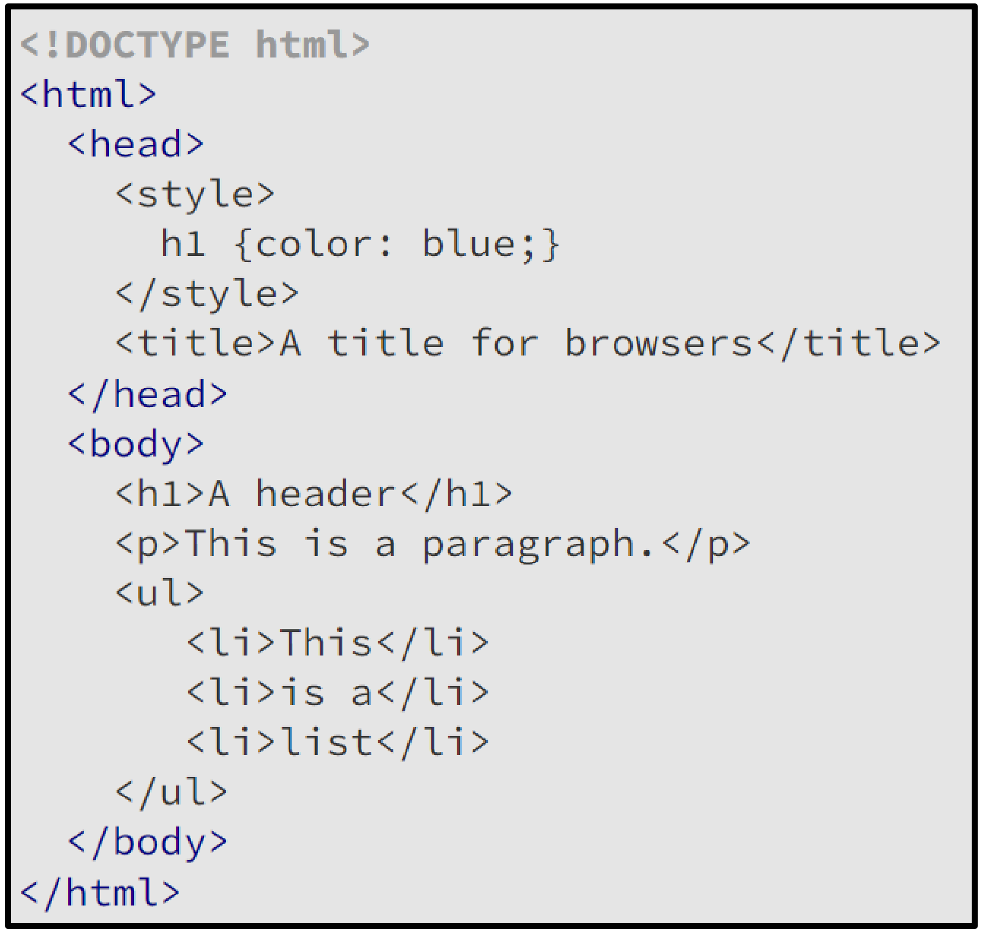
- one compulsory and non-credit seminar course (CSSM 590),
- one compulsory and non-credit thesis course (CSSM 595),
- one compulsory and 3-credit introduction to computational social sciences
- one compulsory and 3-credit programming course (CSSM 502),



```
                 <div class="vc_row wpb_row vc_row-fluid"><div class="
    <ul><li><a href="https://gsssh.ku.edu.tr/en/departments/computational
    In order to graduate from the Computational Social Sciences Master&#8
<p>The list of courses should include:</p>
<ul>
<li>one compulsory and non-credit seminar course (CSSM 590),</li>
<li>one compulsory and non-credit thesis course (CSSM 595),</li>
<li>one compulsory and 3-credit introduction to computational social scie
<li>one compulsory and 3-credit programming course (CSSM 502),</li>
<li>one compulsory and 3-credit theory course (offered by the institute o
<li>minimum two 3-credit electives in computational social sciences and m
</ul>
<p></div><div class="general-under-content">
    </p>
<table width="707">
<tbody>
<tr>
<td width="284"><strong>1. Semester (Fall)</strong></td>
<td width="57"><strong>Credit</strong></td>
<td width="53"><strong>ECTS</strong></td>
```

HTML Basics – Source Code

- HTML stands for *hypertext markup language*: it gives the structure to what is visible to visitors (text, images, links)
- Consists of elements written in between opening and closing tags
- ***html*** holds together the root element; it is also the parent to all other elements.
- ***head*** contains metadata, such as titles and style elements
- ***body*** contains the elements in the main body of pages, such as headers, paragraphs, lists, tables, images
- Most elements have opening and closing **tags**



```html
<!DOCTYPE html>
<html>
  <head>
    <style>
      h1 {color: blue;}
    </style>
    <title>A title for browsers</title>
  </head>
  <body>
    <h1>A header</h1>
    <p>This is a paragraph.</p>
    <ul>
       <li>This</li>
       <li>is a</li>
       <li>list</li>
    </ul>
  </body>
</html>
```

HTML Basics – Source Code

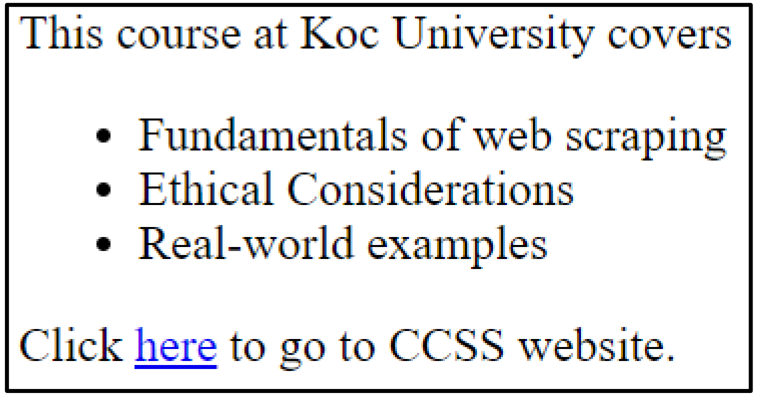
- Most Elements have opening and closing tags: **<p>**This is a paragraph content**</p>**

- Some of the most used tags include:

```
<p>This course at Koç University covers</p>

<ul>
<li>Fundamentals of web scraping</li>
<li>Ethical Considerations</li>
<li>Real-world examples</li>
</ul>

<p>Click <a href="https://ccss.ku.edu.tr/here</a> to go to
CCSS website.</p>
```



This course at Koc University covers

- Fundamentals of web scraping
- Ethical Considerations
- Real-world examples

Click here to go to CCSS website.

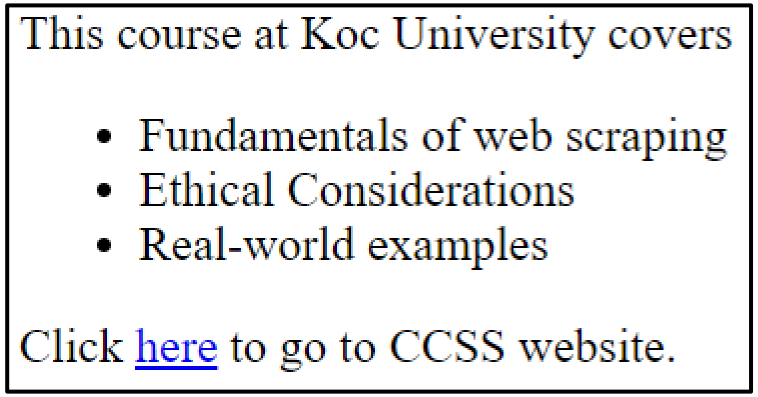
HTML Basics – Source Code

- Elements can have attributes: identifiers that separate from other similar contents or group them together.
- These are either **classes** or **IDs**. They allow us to select / target particular contents.
- They’re only visible in the back-end (i.e., the source code).

```
<p class=“paragraph1”>This course at Koç University
covers</p>

<ul>
<li id=“list_item1”>Fundamentals of web scraping</li>
<li id=“list_item2”> Ethical Considerations</li>
<li id=“list_item3”> Real-world examples</li>
</ul>

<p class=“paragraph2”> Click <a
href="https://ccss.ku.edu.tr/here</a> to go to CCSS
website.</p>
```

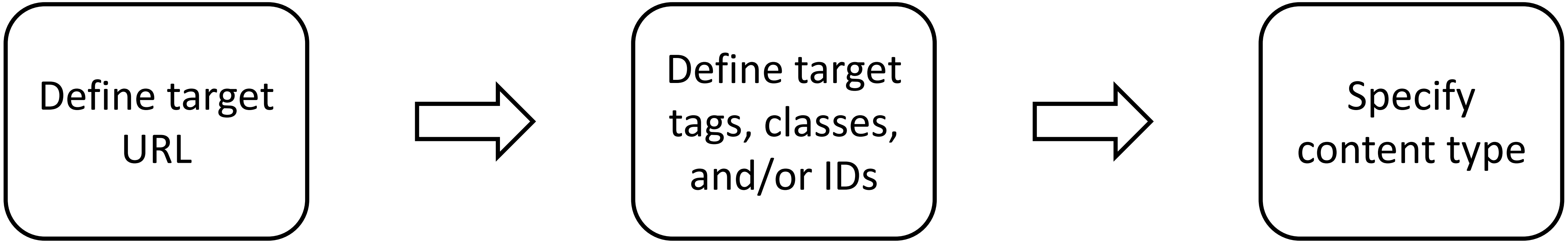


This course at Koc University covers

- Fundamentals of web scraping
- Ethical Considerations
- Real-world examples

Click here to go to CCSS website.

The Scraping Process



Define target URL

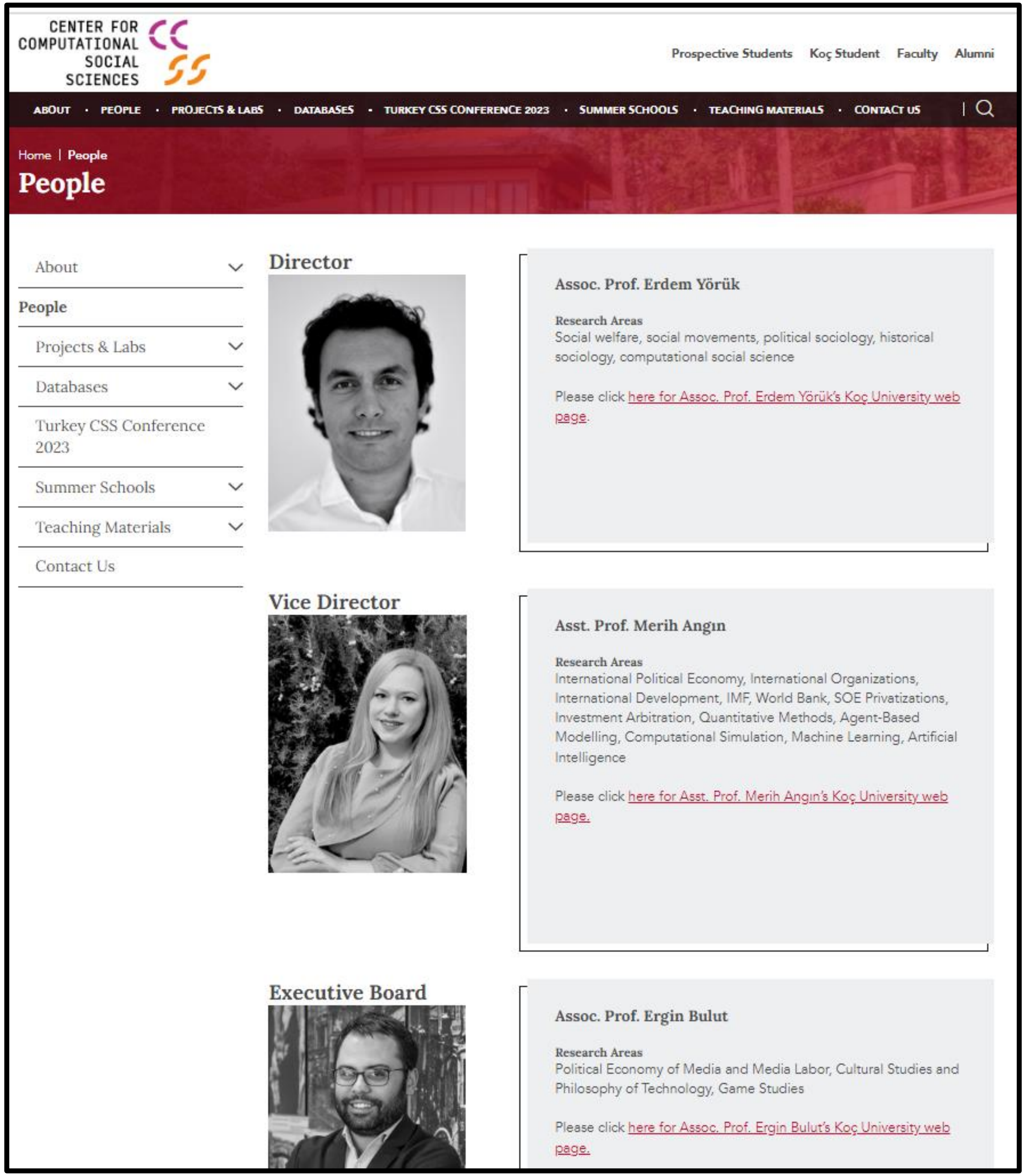
Define target tags, classes, and/or IDs

Specify content type

- **Target URL:** The web page where the content is hosted (eg, the *'people'* page on CCSS website).

- **Target tags, classes, and/or IDs:** individual or a combination of elements that you want to scrape.

- **Content type:** what type of content are you looking to collect (text, tables, links, etc.)?



People

Director

Assoc. Prof. Erdem Yörük

Research Areas
Social welfare, social movements, political sociology, historical sociology, computational social science

Please click here for Assoc. Prof. Erdem Yörük's Koç University web page.

Vice Director

Asst. Prof. Merih Angın

Research Areas
International Political Economy, International Organizations, International Development, IMF, World Bank, SOE Privatizations, Investment Arbitration, Quantitative Methods, Agent-Based Modelling, Computational Simulation, Machine Learning, Artificial Intelligence

Please click here for Asst. Prof. Merih Angın's Koç University web page.

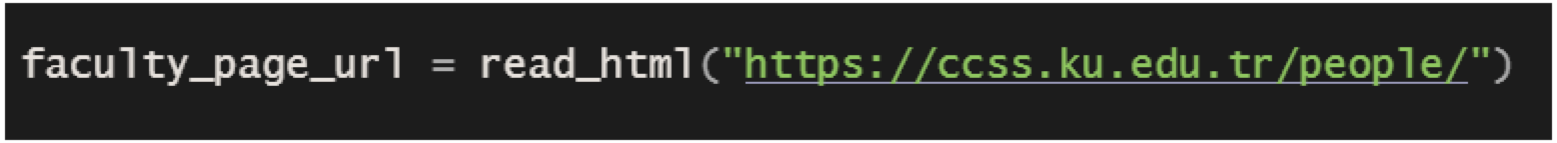
Executive Board

Assoc. Prof. Ergin Bulut

Research Areas
Political Economy of Media and Media Labor, Cultural Studies and Philosophy of Technology, Game Studies

Please click here for Assoc. Prof. Ergin Bulut's Koç University web page.

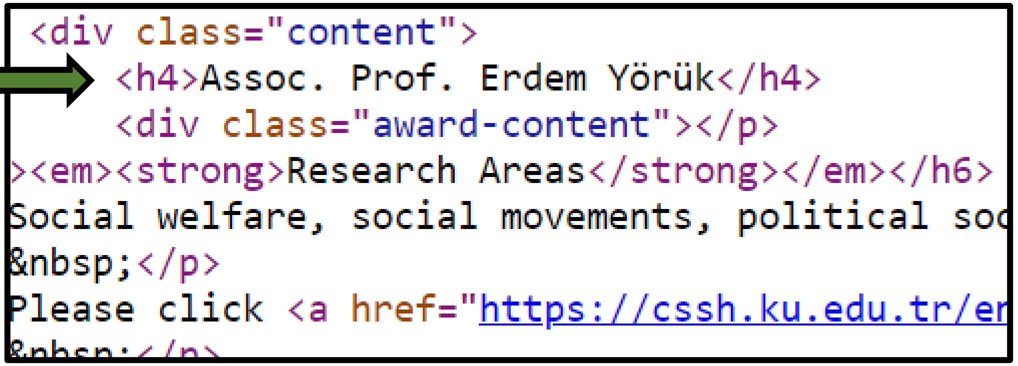
Step 1: Define target URL (read_html function)



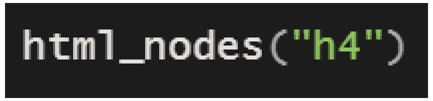
```
faculty_page_url = read_html("https://ccss.ku.edu.tr/people/")
```

Step 2: Look-up and specify the desired attributes and tags (html_nodes function)

Faculties’ names and titles are nested within the **<h4></h4>** tag.

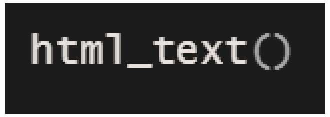


```
 <div class="content">
    <h4>Assoc. Prof. Erdem Yörük</h4>
    <div class="award-content"></p>
><em><strong>Research Areas</strong></em></h6>
Social welfare, social movements, political soc
&nbsp;</p>
Please click <a href="https://cssh.ku.edu.tr/en
&nbsp;</p>
```



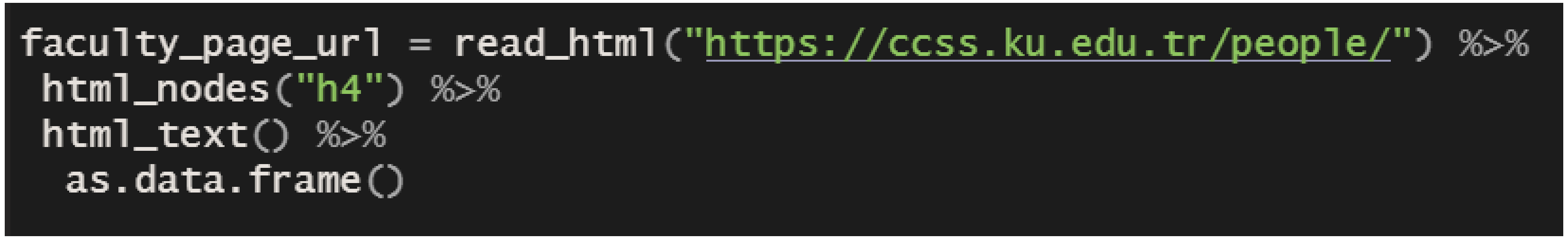
```
html_nodes("h4")
```

Step 3: Specify content type (in this case, it is text)



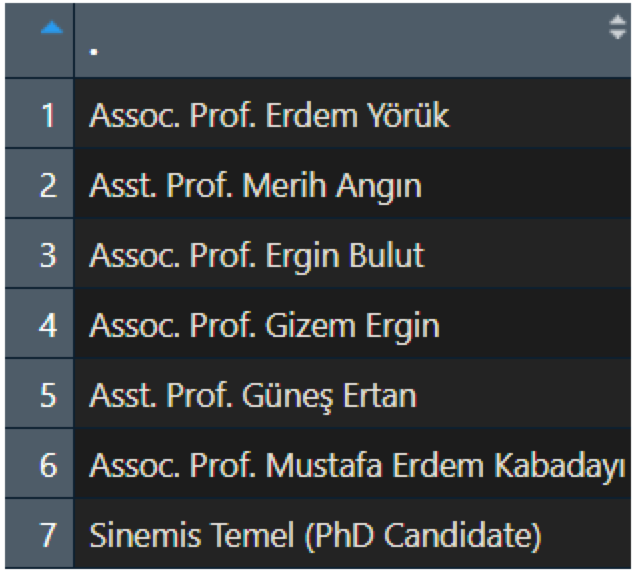
```
html_text()
```

Code:



```
faculty_page_url = read_html("https://ccss.ku.edu.tr/people/") %>%
 html_nodes("h4") %>%
 html_text() %>%
  as.data.frame()
```

Output:

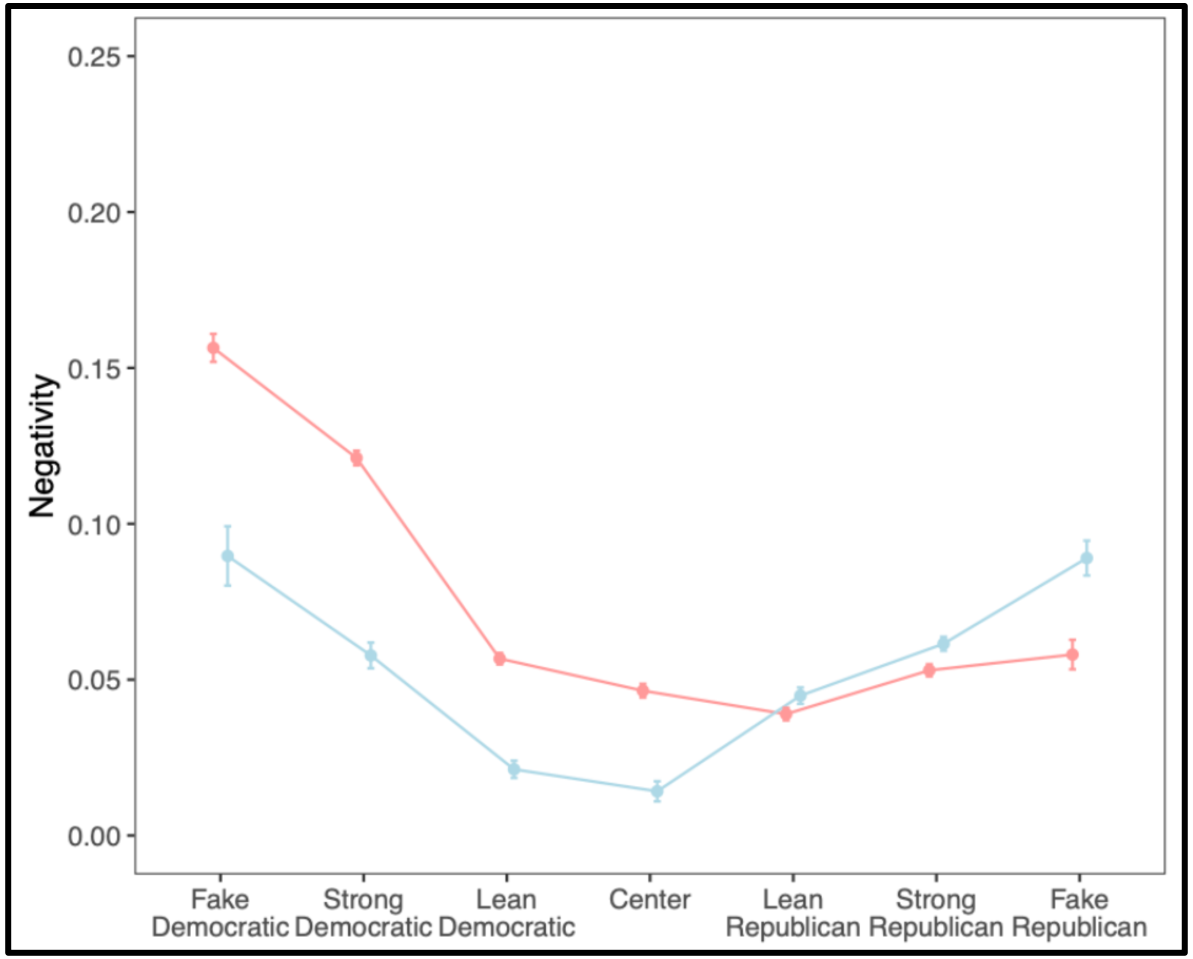


	.
1	Assoc. Prof. Erdem Yörük
2	Asst. Prof. Merih Angın
3	Assoc. Prof. Ergin Bulut
4	Assoc. Prof. Gizem Ergin
5	Asst. Prof. Güneş Ertan
6	Assoc. Prof. Mustafa Erdem Kabadayı
7	Sinemis Temel (PhD Candidate)

How partisan polarization drives the spread of fake news

Mathias Osmundsen, Michael Bang Petersen, and Alexander Bor

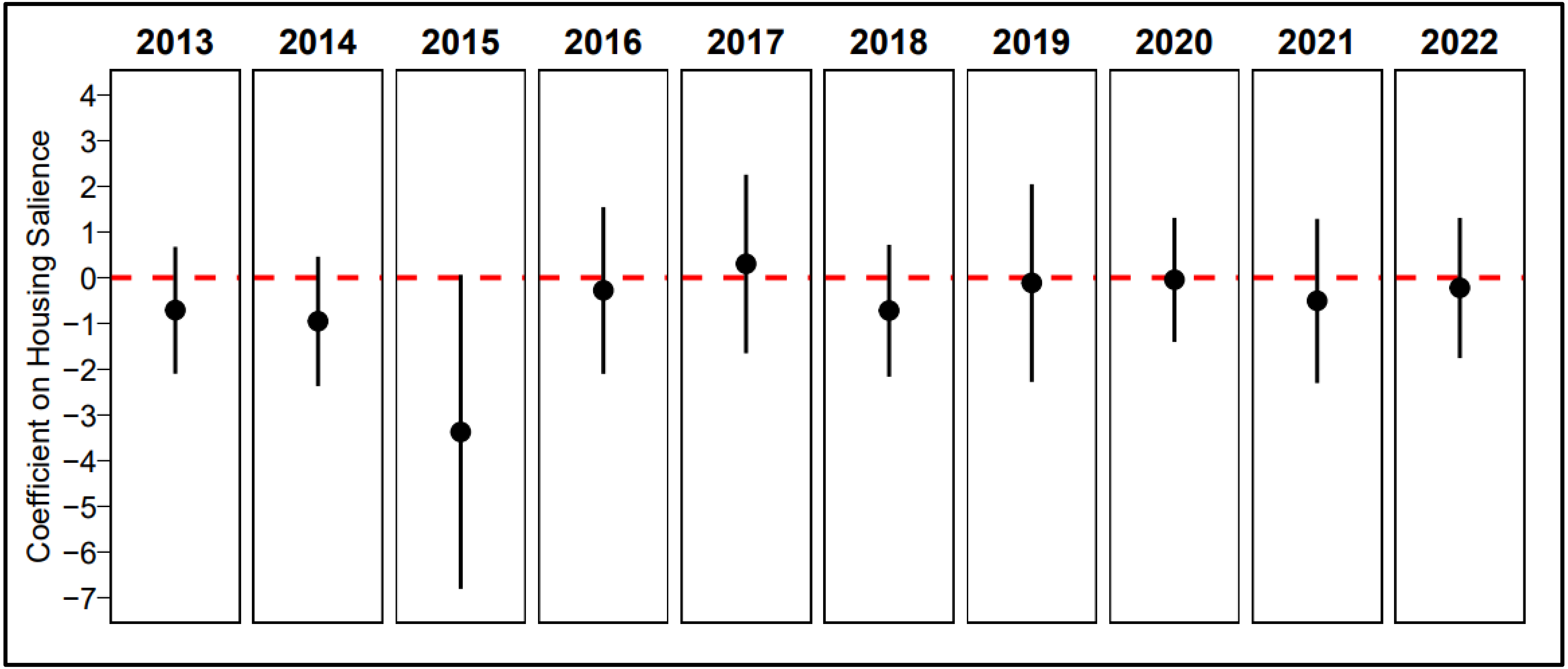
- Scraped over 500,000 news articles headlines shared by social media users
- Goal: analyze negativity trends across partisan groups (in the U.S.)



Legislating Landlords: Private Interests, Issue Emphasis, and Policy Position

Stefan Müller and Jihed Ncib

- We collected 450,773 questions posted by Irish members of Parliament
- Goal: examine whether landlords avoid talking about housing compared to non-landlords


Coefficient on Housing Salience