WEB SCRAPING AND SOCIAL MEDIA SCRAPING

Organisational matters, introduction & legal issues

Maciej Świtała, PhD
Ewa Weychert

Spring 2026

UNIVERSITY OF WARSAW | FACULTY OF ECONOMIC SCIENCES

- Maciej Świtała
- PhD in social sciences in the discipline of economics and finance
- Master of laws
- Research interests: natural language processing, machine learning, empirical legal studies

- Ewa Weychert
- Research interests: demography, machine learning, natural language processing
- Collaborates with LabFam (Interdisciplinary Centre for Labour Market and Family Dynamics) and Florence University

1. Organisational matters
   - Plan of the course
   - Assessment of the course
   - Recommended literature
   - Class materials
   - Programming language to be used
2. Introduction to web scraping
   - Crucial concepts
   - What is web scraping used for?
3. Legal issues
   - Legal compliance in web scraping 101
   - Robots Exclusion Protocol
   - Terms of Service / Terms of Use
   - Applicable legal regulations
   - Intellectual property
   - Confidential and personal data

Organisational matters

Plan of the course

1. Feb 17 - Organisational matters, introduction & legal issues (with MŚ)
2. Feb 24 - Basic HTML navigation (with EW)
3. Mar 10 - Regex (with EW)
4. Mar 17 - Scraping with requests and BeautifulSoup (with EW)
5. Mar 24 - Scraping with Selenium (with MŚ)
6. Mar 31 - Efficient webscraping with Scrapy (with MŚ)

Assessment of the course

- Class attendance: minimum 50% presence required; up to 1 unjustified absence allowed.
- Final grade: **home-taken project** (50% weight) **and its defense** (50% weight).
- Minimum requirement to pass: 50% from each component.
- The project is to be completed as a home assignment.
- Once the project is submited, we arrange its defense, i.e., we meet and ask you questions about it.
- Project submission deadlines are: **May 12** (first take), and **Aug 17** (second take).
- Project defences are scheduled with individual slots in alphabetical order for: **May 26, Jun 9** (first take), and **Aug 31 - Sep 13** (second take). **These are non-negotiable**.

Assessment of the course: project requirements

1. Projects must be completed **individually**.
2. Choose a website to scrape and consult your choice with the instructors before starting, i.e., **send us project proposals in PDF format (up to 1 page) via Moodle WNE, named "name_surname.pdf", until Apr 7**. You must get our acceptance of it.
3. The selected website **must be legal** to scrape.
4. Scrape the website in a way that demonstrates your familiarity with:
   - the HTML structure of websites,
   - Python regular expressions,
   - the following Python libraries: requests, BeautifulSoup, Selenium, Scrapy.

   i.e., **all those elements should be included in the final project**.
5. The final output should be a **dataframe containing the structured data you scraped**.
6. Provide **a report** describing all your actions, preferably in Jupyter Notebook.
7. Submit your project via Moodle WNE as a ZIP file named exactly **"name_surname.zip"**.

Assessment of the course: project requirements

The ZIP file must contain:

- your Python code,
- the dataset (if it is small enough to include directly),
- a **"README.txt"** file containing:
  - your name and surname,
  - a list of all files and the order in which they should be run,
  - a link to Google Drive with your scraped data (if the dataset is too large to include it directly).
- a **"requirements.txt"** file containing a list of all packages used,
- a **"legal_proof.txt"** file including a written proof that the webpage is legal to scrape.

Assessment of the course: project requirements

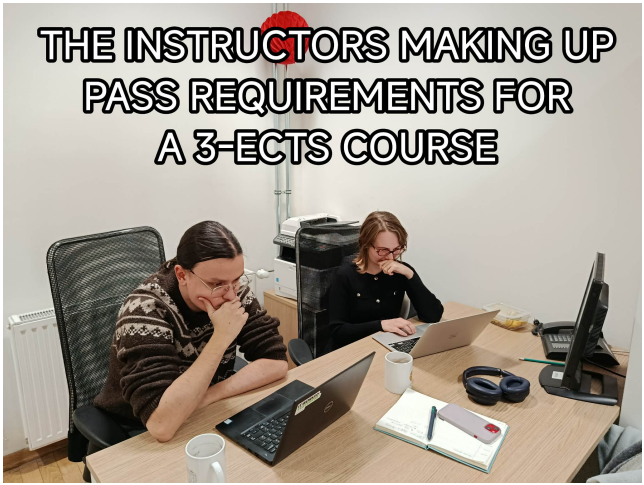
- Evaluation will be based on:
  - technical **correctness** in applying the tools,
  - **efficiency** of the proposed solution,
  - **preparation of the data** for further analysis,
  - structure, i.e., **readability, and clarity** of the report.
- Additionally, **the complexity of your project** will be evaluated in comparison to projects completed by other students, i.e., not only correctness and functionality are assessed, but also:
  - the scope of work,
  - the difficulty of the implemented solutions,
  - the overall technical advancement of the project.
- Therefore, your project should not only meet the functional requirements but also demonstrate **an appropriate level of ambition and independent work**.

Assessment of the course: project defence

- Each student will be required to defend their project.
- The purpose of the defence is to verify that the project was completed independently and that the student fully understands the tools, and code used.
- **No presentation** needs to be prepared.
- During the defense:
  - your code will be displayed,
  - instructors will ask **three questions** possibly related to:
    - your project implementation and code,
    - general Python programming knowledge,
    - web scraping concepts and techniques used in your projects.
- You must be able to clearly explain how your solution works and justify the choices you made.
- **The evaluation will be based solely on your responses to questions we ask**.

Assessment of the course: major remark

- Each incompliance with technical requirements, e.g., not getting our approval for the project proposal, not submitting the project via Moodle WNE, or naming files differently from our guidelines, will lower your grade by 0.5.



Recommended literature

- R. Mitchell (2018). Web Scraping with Python: Collecting Data from the Modern Web. 2nd Edition. O'Reilly Media.

Class materials

Shared via the elearning.wne.uw.edu.pl

Contact information

ms.switala@uw.edu.pl
e.weychert@uw.edu.pl

Programming language to be used

- Python programming language is planned to be used.
- A vast majority of you already knows (or at least should know) Python.
- Python is free, efficient and fast (relatively).
- It is moderately preffered by the employers.

Introduction to web scraping

Motivation

- The internet contains vast amounts of data.
- Using programming tools, we can **automatically collect data**, and optionally also update it in real-time.
- Ultimately, we want to conduct further analyses based on the collected data, i.e., build predictive models, recommendation systems, etc.
- Web scraping is often just the very first, technical step toward something far more interesting.

What is web scraping?

- Constructing a precise definition based on the literature, web scraping can be described as **the process of automatically collecting and acquiring data available on the Internet**, using methods other than APIs, which are pre-built infrastructures for data retrieval, **and transforming it into a structured form**, enabling further processing and analysis.
- Thus, a (web) scraper is a program that obtains data from multiple websites of the same type and processes it for further analysis.

Related concepts: crawling

- Intuitively, a crawler (robot, spider, fish, worm) is a program that simulates the behavior of a website user.
- Therefore: **crawling simulates user behavior** (clicking buttons, logging in, copying, pasting, etc.), while web scraping focuses on retrieving and processing data.
- That is: **first comes crawling, then web scraping**.
- Another interpretation: web scraping is a broader concept than crawling and includes it.
- In practice, the difference between these concepts is often ignored, and the two terms are used interchangeably.

Related concepts: web mining, bot

- Web mining refers to broadly understood **"data mining" from the Internet**.
- It seems that this is a broader concept than web scraping, in particular it also includes using APIs.
- A bot is **a program that performs repetitive tasks**.
- A bot is a broader concept than a scraper and a crawler.

What is web scraping used for?

It is, of course, used to collect data for further analysis, yet sometimes web scraping itself is an important part of the final product. Especially in the case of:

- automated website testing,
- competitor research,
- real-time sentiment and opinion analysis,
- dynamic pricing,
- brand protection,
- recommendation systems.

See more: https://research.aimultiple.com/web-scraping-applications/.

What will we automate?

- Opening a browser.
- Gaining access to websites.
- Clicking buttons.
- Logging in/out.
- Retrieving, copying, and saving data of interest.

What problems will we encounter?

- Web scraping may violate the law. Always ensure that your actions **comply with applicable legal regulations**.
- The scraper should be **efficient**, i.e., it should collect as much data as possible in the shortest time as possible, while minimizing errors.
- Even when scraping is legal, website administrators often implement measures to make it more difficult in order to prevent performance degradation or access issues. Therefore, we must **scrape in a way that minimizes detection to avoids bans or CAPTCHAs**.

Legal issues

Is web scraping legal?

- **In general - yes**: scraping publicly available content is legal.
- However, **the interests of entities providing access to websites** must be taken into account.
- In addition, the law regulates **confidential data, personal data, and intellectual property rights**.

How to check whether data can be scraped from a specific source

1. Check the contents of the **Robots Exclusion Protocol**, i.e., robots.txt file.
2. Review the website's **Terms of Service / Terms of Use**.
3. If the above do not provide a clear answer, analyse **applicable legal regulations**.

How to check whether data can be scraped from a specific source: step 1 of 3

- It is recommended to begin by examining the **robots.txt** file.
- This file is based on a standard introduced in 1994, known as the **Robots Exclusion Protocol**.
- It provides **non-legally binding** guidelines for automated agents ("robots").
- Most websites provide such a file, typically accessible at: `.../robots.txt`.
- Common directives include: **User-agent, Disallow, Allow, *, /, Crawl-delay**.

Example 1

- Let us consider: `facebook.com/robots.txt`.
- At the beginning of the file, Facebook states: *"Collection of data on Facebook through automated means is prohibited unless you have express written permission from Facebook"*.
- At the end of the document, the directive **"User-agent: * Disallow: /"** appears, meaning all automated activity is prohibited.
- Exceptions are provided above for specific bots (e.g., Applebot, Googlebot), with clearly defined permissions.

Example 2

- Consider: reddit.com/robots.txt.
- The file does not contain an explicit warning message at the beginning.
- The directive **"User-agent: *"** indicates that the rules apply to all bots, except those explicitly listed.
- Scraping appears to be generally permitted as we see **"User-agent: *"**.

Example 3

- Consider: x.com/robots.txt.
- The document does not include an explicit introductory warning.
- The directive **"User-agent: * Disallow: /"** indicates that all automated activity is prohibited.
- There are also specific regulations regarding Googlebot, facebookexternalhit, Google-Extended, FacebookBot, DiscordBot, and BingBot; for some all activities are prohibited with **"Disallow: *"**.

How to check whether data can be scraped from a specific source: step 2 of 3

- **The provisions of the robots.txt file are not legally binding**; it is just a good practice to follow it.
- To assess legality, it is necessary to review the website's **Terms of Service / Terms of Use**.
- From a legal perspective, these terms constitute **a contract between the service provider and the user**.
- The user becomes a party to the contract upon using the website.
- The contract may prohibit scraping and specify consequences for prohibited actions, including:
  - temporary restrictions, rate limiting,
  - suspension of the account,
  - permanent account termination (ban),
  - bans extended to related accounts (same IP, device, cookies),
  - liquidated damages.
- If there are no consequences specified, breaching of contract implies its termination, i.e., deleting the account, ban.
- Breach of contract that results in **damage may give rise to liability for compensation, even if such consequences are not specified in the**

Example 4

- Consider the website `imdb.com`. Its `robots.txt` file suggests that scraping is allowed (yet with many exceptions).
- However, the `Terms of Service` explicitly state: **"You may not use data mining, robots, screen scraping, or similar data gathering and extraction tools on this site, except with our express written consent as noted below"**.
- Simply put: the Terms of Service are legally-binding, not the robots.txt file.

Example 5

- Consider the website `wattpad.com`. Its `robots.txt` file suggests that scraping is allowed (yet with many exceptions).
- However, the `Terms of Service` explicitly state: **"Don't scrape Wattpad"**.
- Again: the Terms of Service are legally-binding, not the robots.txt file.

How to check whether data can be scraped from a specific source: step 3 of 3

- Even if the Terms of Service are absent or incomplete, **unlawful conduct that causes a material damage may result in civil liability**.
- Material damage (property damage) refers to any loss, destruction, or impairment of physical property that reduces its value or usability. It can include damage to buildings, vehicles, equipment, or other tangible assets.
- Also, criminal liability may arise in certain factual states, e.g., under Chapter XXXIII of the Polish Penal Code, i.e., crimes against information protection.

Example 6

- Let us consider a program that crawls the website of the Warsaw City Hall.
- The website does not provide a `robots.txt` file.
- There are no formal Terms of Use, primarily as the website is operated by a public authority.
- Can automated appointment scheduling for personal use be considered lawful?

How to check whether data can be scraped from a specific source: step 3 of 3 - cont'd

- **Intellectual property** is subject to special legal protection.
- In general, previously published **works** may be used without the author's consent **for personal use only**.
- Simply put, **educational and research purposes**, as long as they are non-commercial and one cites the author, are also allowed.
- Hereby a **work** is defined as any manifestation of creative activity with an individual character.
- Commercial use requires the author's explicit consent.

How to check whether data can be scraped from a specific source: step 3 of 3 - cont'd

- **Confidential and personal data** are subject to special protection.
- Confidential data include non-public information with economic value.
- Personal data concern identified or identifiable natural persons.
- In the EEA, personal data processing is regulated by the **GDPR**.
- The GDPR does not apply to purely personal, non-commercial activities.

How to check whether data can be scraped from a specific source: summary

1. The **robots.txt** files provide guidelines, not legally binding rules.
2. The **Terms of Service** may prohibit scraping and impose contractual consequences.
3. Violations of general legal norms may result in **liability for material damage**, eventually compensation.
4. Scraping, in specific factual states, may also lead to **criminal liability**.
5. Particular caution is required when dealing with **copyrighted works, confidential data, and personal data**.

Thank you for your attention!

Maciej Świtała, PhD
ms.switala@uw.edu.pl
Ewa Weychert
e.weychert@uw.edu.pl