Programming and web data collection

ECTS : 3

Description du contenu de l'enseignement :

This course covers essential Python programming techniques for web data collection in applied economic analysis. Students will learn practical methods to extract structured data from online sources, starting with basics such as HTML/CSS, HTTP requests, XPath, CSS selectors, browser emulation, and public/private APIs (World Bank, INSEE, IMF).

Advanced topics include hidden APIs, overcoming technical obstacles (session management, blocking points), and large-scale data extraction. Students will gain expertise using libraries like requests, BeautifulSoup, and pandas for JSON/XML handling, data cleaning, and pipeline creation.

The course also emphasizes ethics, legal compliance, privacy, and responsible data use. Practical exercises and real-world examples will enable students to develop robust solutions for collecting and analyzing economic data from the web.

Compétence à acquérir :

Course Objectives:

- Write structured and reusable Python code for data tasks.
- Interact with APIs and process JSON/XML data structures.
- Understand HTML structure and use scraping tools like BeautifulSoup.
- Automate web data collection while following ethical standards.
- Clean, structure, and store collected data for analysis.
- Identify and navigate common technical challenges in web scraping.
- Implement browser emulation techniques for complex data collection scenarios.
- Build reproducible data pipelines to facilitate economic research and analysis.
- Evaluate legal constraints and ethical implications of web data extraction.

Targeted competencies:

- Develop robust Python programming skills tailored to data collection and analysis.
- Effectively utilize REST APIs and parse structured web data (JSON/XML).
- Extract data reliably from static web pages using scraping tools such as BeautifulSoup.
- Efficiently clean and transform datasets using pandas and regular expressions (regex).
- Design and document reproducible pipelines for systematic data acquisition and analysis

Mode de contrôle des connaissances :

The assessment will consist in written exam and an oral presentation of a project made in groups.

Bibliographie, lectures recommandées :

Python and Scraping

- https://developers.google.com/edu/python/introduction
- https://arxiv.org/abs/2211.04630
- https://realpython.com/python-web-scraping-practical-introduction

Ethical, Legal, and Practical Considerations

- https://arxiv.org/abs/2410.23432
- https://www.cnil.fr/fr/focus-interet-legitime-collecte-par-moissonnage
- https://www.captaincontrat.com/protection-des-creations/cgv-cgu-cga/web-scraping-est-ce-legal-me-marcotte
- https://fr.wikipedia.org/wiki/Donn%C3%A9es_ouvertes_en_France