

MODULE SPECIFICATION • ELECTIVE

Web Scraping and end-to-end data pipelines

The purpose of this course is to provide an introduction to web scraping. It covers basic techniques of web scraping, reviews common libraries and frameworks for web scraping in Python, extraction from HTML and XML, and discusses more advanced techniques. This course also covers the basics of data engineering, including data ingestion, cleaning and transformation, as well as data storage and retrieval.

10 CREDITS	15h CONTACT HOURS	85h PRIVATE STUDY	100h TOTAL STUDY
---------------	----------------------	----------------------	---------------------

COURSE CONVENOR

NAME Régis Amichia

OFFICE By appointment — please email to arrange a convenient time

COURSE STRUCTURE

Econ 624 is a 10-credit course and students are expected to input approximately 100 hours of study. The total number of contact hours is 15 hours, leaving 85 hours for private study. Delivery is by lectures — 15 hours across the first three weeks of term, comprising 10 hours of lectures and 5 hours of tutorials. Optional clinics run on the last day of the course.

During private study, balance the course material (the primary source) against the recommended textbooks: think critically about how these fit into the body of knowledge, perform the exercises, complete coursework and revise. You can expect to perform well only by working consistently through the year.

MODULE PROGRAMME

LECT.	TOPIC	SESSION CONTENT
01	Introduction to Web Scraping	- Definition and purpose - Legality and ethical considerations - Internet in a nutshell - Basics of HTML - Introduction to BeautifulSoup - Reading documents
02	Advanced web scraping	- Data extraction forms and logins - Introduction to Selenium - Dynamic websites and JavaScript - Web scraping best practices
03	Introduction to Data Engineering	- What is data engineering - Data engineering in organizations - Main data engineering architecture - The process of data transformation
04	Apache Airflow and data orchestration	- Introduction to orchestrators and Apache Airflow - Airflow main concepts - Create and run automated data pipelines - Monitor and debug pipelines

KEY SKILLS

By the end of this course, students should have some knowledge and understanding of:

- **Data engineering basics.** Fundamental understanding of data engineering principles.
- **Data pipeline setup.** Ability to set up data pipelines for efficient data processing.
- **Web scraping techniques.** Skills in collecting data from the web using scraping methods.
- **Tool proficiency.** Familiarity with tools necessary for data engineering tasks.
- **Data collection from the web.** Competence in extracting data from web sources.
- **Practical application.** Applying data engineering skills to real-world scenarios.
- **Data processing skills.** Handling and processing data efficiently within pipelines.
- **Understanding tools for data engineering.** Knowledge of tools essential for data engineering tasks.

DESIRED OUTCOMES

By the end of this course, students should be able to:

- Understand the basics of web scraping and its applications
- Extract data from HTML and XML using web scraping techniques
- Use common libraries and frameworks for web scraping in Python, such as BeautifulSoup and others
- Handle advanced web scraping challenges, such as dealing with dynamic websites and avoiding detection
- Communicate effectively about web scraping techniques and their applications
- Understand the basics of data engineering and the role of end-to-end data pipelines in data analytics



- Design and implement data pipelines for data ingestion, cleaning and transformation
- Use common tools and technologies for building data pipelines, such as Apache Airflow
- Monitor and troubleshoot data pipelines for performance and reliability
- Communicate effectively about data engineering concepts and techniques

ASSESSMENT

WEIGHTING	100% coursework
ISSUED	End of week 6
DUE	End of week 10 — four weeks for completion
SUBMISSION	Electronically via the Moodle site for this course
FEEDBACK	Returned within 4 weeks of the deadline, as marker's notes appended to the PDF of your coursework

Marking criteria and penalties are set out in the Economics Undergraduate Handbook and the general course information paper, available via the Current Student page of the university website under Academic Regulations.

COURSE TEXT & RECOMMENDED READING

Book chapters and online material. Copies of the lecture slides are made available on the course web pages, together with additional material and course announcements.

- Mitchell, R. Web Scraping with Python
- Harenslak, B. P. and de Ruiter, J. R. Data Pipelines with Apache Airflow
- Reis, J. and Housley, M. Fundamentals of Data Engineering