Lancaster University Management School

timberlake
statistics | econometrics | forecasting

# Econ624 : Web Scraping and end-to-end data pipelines

## COURSE AIMS & OBJECTIVES, KEY SKILLS AND LEARNING OUTCOMES

**Course Aims & Objectives:** This course is 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.

**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 ingestion, such as Apache Airflow;
- Monitor and troubleshoot data pipelines for performance and reliability;
- Communicate effectively about data engineering concepts and techniques.

## COURSE STRUCTURE

Econ 624 is a 10 credits course and therefore students are expected to input approximately 100 hours of study into the course. The total number of contact hours on Econ 624 is 15 hours. This leaves 85 hours for private study. Course Delivery comes in the form of Lectures with 15 hours delivered over the first 3 weeks of the term (10 hours of lectures and 5 hours of tutorials). There will be optional clinics on the last day of the course.

During your private study you should strike a balance between reading the course material (which is the primary source of information) and the recommended textbooks, thinking critically about how these fit in to the body of knowledge on the subject and about how our level of knowledge can be improved, performing exercises, completing coursework and revising for examinations. You can expect to perform well on this course only if you work consistently through the year.

**Prior to enrolment on the module, the student must have successfully completed**
**Econ620**
**Econ621**

## COURSE CONVENOR

Régis Amichia

## LECTURERS CONTACT INFORMATION (Including Office Hours)

Regis Amichia

email: regis@foxintelligence.io

Thomas Pical

email: tpical@equancy.com

## COURSEWORK ASSESSMENT

The CWA mark will be calculated as 100% coursework. The coursework will be assigned at the end of the course

The coursework will be delivered to students at the end of week 6 of each term and is due for submission at the end of week 10 of the term, allowing students 4 weeks for completion.

Coursework must be submitted electronically through the Moodle site for this course:

## FEEDBACK ON COURSEWORK:

The coursework will be marked and returned to students within 4 weeks of the submission deadline. Feedback will consist of marker's notes appended to the pdf of your coursework.

## MARKING CRITERIA AND PENALTIES

Marking criteria can be found in the Economics Undergraduate Handbook and the general course information paper. An electronic copy of this can be found via the Current Student page of the university website then follow the Academic Regulations link
https://gap.lancs.ac.uk/ASQ/QAE/MARP/Documents/UG-Assess-Regs.pdf

## FINAL MARK INFORMATION

This course is assessed 100% by means of coursework. The final mark is the average of the marks obtained in the two pieces of coursework.

## COURSE TEXT AND RECOMMENDED READING

Lecture notes and Lecture slides.

Web Scraping with Python" by Ryan Mitchell

Data Pipelines with Apache Airflow by Bas P. Harenslak, Julian Rutger de Ruiter

Fundamentals of Data Engineering by Joe Reis, Matt Housley

## MARKING CRITERIA AND PENALTIES

Marking criteria can be found in the Economics Undergraduate Handbook and the general course information paper. An electronic copy of this can be found via the Current Student page of the university website then follow the Academic Regulations link
https://gap.lancs.ac.uk/ASQ/QAE/MARP/Documents/UG-Assess-Regs.pdf

## COURSE OUTLINE/LECTURE SCHEDULE

**Lecture 1: Introduction to Web Scraping**

- Definition and purpose
- Legality and Ethical Considerations
- Internet in a nutshell

- Basics of HTML
- Introduction to Beautifulsoup

**Lecture 2: Advanced web scraping**

- Reading documents
- Data extraction forms and logins
- Introduction to Selenium
- Dynamic websites and JavaScript
- Web scraping best practices

**Lecture 3: Introduction to Data Engineering**

- What is data engineering
- Data engineering in organizations
- Main data engineering architecture
- The process of data transformation

**Lecture 4: Apache Airflow and data orchestration**

- Introduction to orchestrators and Apache Airflow
- Airflow main concepts
- Create and run automated data pipelines
- Monitor and debugging pipelines