ESCUELA TÉCNICA SUPERIOR DE INGENIERÍA

DTC-MBD-521 Data Acquisition and Transformation

SEMESTER: Winter

CREDITS: 30 hours

LANGUAGE: Spanish/English

DEGREES: Master in Big Data Technologies and Advanced Analytics

Course overview

This course is an introduction to the fundamental principles of data acquisition and transformation. The aim of this subject is to provide to the students the techniques, methodologies and tools to successfully obtain, clean, correct, standardize and transform data from different information sources.

Prerequisites

Basic knowledge of Python is required.

Course contents

Theory:

1. Intro. ¿Para qué me sirve? Casos de uso reales.
2. Requests + Postman.
3. Caso de uso API (skyscanner, ryanair, ...).
4. Formatos de respuesta: nested dict, array of dicts, array of arrays. JSON vs CSV.
5. Estructura de XML y HTML. Intro a BS4.
6. Parsear XML y HTML con BS4. Casos básicos (Wikipedia).
7. Parsear XML y HTML con BS4. Casos complejos (loops del INE).
8. Interactuar con JS. Selenium + chromedriver.
9. Interceptar peticiones JS desde el navegador.
10. Selenium + chromedriver. Casos básicos (OCDE).
11. Selenium + chromedriver. Casos complejos (Seace). Wireshark.
12. Desestructurado a estructurado. Leyendo PDFs con Pytesseract.
13. Proyecto completo: spider + scraper + parser.
14. Logueo de eventos y gestión de errores en el scrapeo.
15. Despliegue sencillo del proyecto en una VM.
16. Resumen y guías para seguir aprendiendo.

Textbook

While we will not follow a textbook, we find the following books quite remarkable in their central topics.

- **Ryan Mitchel, (2018),** *Web scraping with Python. 2nd Edition. O'Reilly Media.*

Grading

The following conditions must be accomplished to pass the course:

- A minimum overall grade of at least 5 over 10.
- A minimum grade in the final test of 5 over 10.

The overall grade is obtained as follows:

- Final test accounts for 30% of the final grade if the grade in this exam is at least 5.
- Laboratory session work (in class and homework) accounts for 70% of the final grade.