FICHA TÉCNICA DE LA ASIGNATURA

Datos de la asignatura	
Nombre completo	Adquisición y transformación de datos
Código	DTC-MBD-521
Título	Máster en Big Data. Tecnología y Analítica Avanzada/Master in Big Data Technologies and Advanced Analytics
Impartido en	Máster en Big Data. Tec. y Analítica Avanzada/Master in Big Data Technologies and Advanced Analytics [Primer Curso]
Créditos	3,0 ECTS
Carácter	Obligatoria
Departamento / Área	Departamento de Telemática y Computación

Datos del profesorado

Profesor

Nombre	Daniel Jadraque de Soria
Departamento / Área	Departamento de Telemática y Computación
Correo electrónico	djadraque@icai.comillas.edu

DATOS ESPECÍFICOS DE LA ASIGNATURA

Contextualización de la asignatura

Aportación al perfil profesional de la titulación

The quality of any statistical model cannot be greater than the quality of the data it is based on. As a result, gaining the ability to gather and cleanse data should be one of the first priorities of any data scientist student, and that is exactly what this course is all about.

Prerequisitos

A basic knowledge of Python is strictly required.

Competencias - Objetivos

Competencias

Competences[1]

General competences

CG1. Have acquired advanced knowledge and demonstrated, in a research and technological or highly specialized context, a detailed and well-founded understanding of the theoretical and practical aspects, as well as of the work methodology in one or more fields of study.

Haber adquirido conocimientos avanzados y demostrado, en un contexto de investigación científica y tecnológica o

	altamente especializado, una comprensión detallada y fundamentada de los aspectos teóricos y prácticos y de la metodología de trabajo en uno o más campos de estudio.
CG2.	Know how to apply and integrate their knowledge, understanding, scientific rationale, and problem-solving skills to new and imprecisely defined environments, including highly specialized multidisciplinary research and professional contexts. *Saber aplicar e integrar sus conocimientos, la comprensión de estos, su fundamentación científica y sus capacidades de resolución de problemas en entornos nuevos y definidos de forma imprecisa, incluyendo contextos de carácter multidisciplinar tanto investigadores como profesionales altamente especializados.*
CG3.	Know how to evaluate and select the appropriate scientific theory and the precise methodology of their fields of study in order to formulate judgements based on incomplete or limited information, including, when necessary and pertinent, a discussion on the social or ethical responsibility linked to the solution proposed in each case. *Saber evaluar y seleccionar la teoría científica adecuada y la metodología precisa de sus campos de estudio para formular juicios a partir de información incompleta o limitada incluyendo, cuando sea preciso y pertinente, una reflexión sobre la responsabilidad social o ética ligada a la solución que se proponga en cada caso.*
CG4.	Be able to predict and control the evolution of complex situations through the development of new and innovative work methodologies adapted to the scientific/research, technological or specific professional field, in general multidisciplinary, in which they develop their activity. *Ser capaces de predecir y controlar la evolución de situaciones complejas mediante el desarrollo de nuevas e innovadoras metodologías de trabajo adaptadas al ámbito científico/investigador, tecnológico o profesional concreto, en general multidisciplinar, en el que se desarrolle su actividad.*
CG5.	Be able to transmit in a clear and unambiguous manner, to specialist and non-specialist audiences, results from scientific and technological research or state-of-the-art innovation, as well as the most relevant foundations that support them. *Saber transmitir de un modo claro y sin ambigüedades, a un público especializado o no, resultados procedentes de la investigación científica y tecnológica o del ámbito de la innovación más avanzada, así como los fundamentos más relevantes sobre los que se sustentan.*
CG6.	Have developed sufficient autonomy to participate in research projects and scientific or technological collaborations within their thematic area, in interdisciplinary contexts and, where appropriate, with a high knowledge transfer component. *Haber desarrollado la autonomía suficiente para participar en proyectos de investigación y colaboraciones científicas o tecnológicas dentro de su ámbito temático, en contextos interdisciplinares y, en su caso, con una alta componente de transferencia del conocimiento.*
CG7.	Being able to take responsibility for their own professional development and their specialization in one or more fields of study. *Ser capaces de asumir la responsabilidad de su propio desarrollo profesional y de su especialización en uno o más campos de estudio.*
Specific competences	

CE5. Know the techniques used to extract information from large datasets, as well as the different platforms, tools, and languages that make it possible.

Conocer las técnicas para extraer información de grandes conjuntos de datos, así como las diferentes plataformas, herramientas y lenguajes que lo hacen posible.

[1] Competences in English are a free translation of the official Spanish version.

Resultados de Aprendizaje

Learning outcomes

By the end of the course students should:

RA1.	Understand how a browser works behind the scenes.
RA2	Understand what an API is and what is it use for.
RA3	Automate data download from both public and private APIs.
RA4	Automate data downloads from both static and interactive websites.
RA5	Structure unstructured data.
RA7	Understand the scalability issues behind any large scraping operation.
RA8	Gain some agility cleansing and wrangling data.

BLOQUES TEMÁTICOS Y CONTENIDOS

Contenidos – Bloques Temáticos

Theory

Unit 0. Introduction

1. Why is this useful?

Unit 1. Postman + HTTP Request

1. What is an HTTP request?
2. Why is an HTTP request needed?
3. What is an API?
4. What is an endpoint?
5. API and endpoint examples
6. Types of HTTP requests
7. How to make a HTTP request in Postman
8. What are query strings and what are they use for?

9. HTTP request status codes
10. HTTP requests with Python requests

Unit 2. API + Python variable types

1. Python variable types
2. String concatenation
3. Analysis of practice API responses: dict and arrays
4. Operations with (nested) arrays
5. Operations with (nested) dictionaries
6. Dealing with JSON responses in Python Requests
7. HTTP request headers description
8. HTTP request analysis of usual headers, meaning and use
9. HTTP request data encoding

Unit 3. Response formats (CSV, JSON, HTML, XML) + Parsing with bs4

1. Understanding client-server requests
2. Intercepting client-server requests in absence of open API
3. CSV format
4. How to write, update and read a CSV file
5. JSON format
6. How to write, update and read a JSON file
7. Best practices for correct data formatting
8. Web intro: html, css and js
9. HTML main tags
10. HTML main attributes
11. XML format
12. Intro to HTML/XML parsing with bs4

Unit 4. Web Scraping

1. Intro to web scraping
2. How to organize a data team
3. How to organize a web scraping project
4. Web scraping vs Brute force
5. Spider phase: description and goals
6. Crawler phase: description and goals
7. Parser phase: description and goals
8. Devops: escalating infrastructure needs
9. DDoS attacks

Unit 5. Interacting with JS

1. Python repo structure
2. Web scraping: dealing with exceptions.
3. Exception logging
4. Exception severity levels
5. Web scraping legal issues.
6. Interacting with JS

7. Selenium basics
8. Chromedriver

Unit 6. More advanced Selenium + Parser phase

1. Concurrent requests: dimensioning
2. Web scraping tech issues
3. Headless Chromedriver
4. IP rotation and proxies
5. Selenium differences vs bs4
6. Captchas and ReCaptchas
7. Unicode
8. Removing non ascii characters
9. Parsing strings: python vanilla and regex
10. OCR for PDF parsing
11. Parsing PDF with Python

Laboratory (Practice topics are tentative and may be replaced by others of similar nature.)

Practice 1. Postman request to Google Places API

Simple practice as ice breaker. Manually input data into Postman to collect a dataset the student is already used to work with in real life: Google Maps data. This practices forces them to read a private API documentation, to understand how to authenticate an API call and to address how easily billing can escalate out of control in automated environments.

Practice 2. AEMET API

The main goal of the practice is to learn how to retrieve data from an open API using an API key. Besides, it shows an example of an API of a public institution to bring it into comparison with private APIs in later practices.

Practice 3. Linkedin

Open question regarding the status code received after querying Linkedin, to validate the student understanding on what's going on in an HTTP behind the scenes.

Practice 4. Dictionary parsing

The goal is to address lack of knowledge in basic Python operations as soon as possible, before getting into more complicated areas. This practice asks the student to retrieve a specific data point from an API that returns a nested dictionary as response

Practice 5. Twitter API

With previous APIs all we manage to do was to grab data. This time we post data to the internet using Twitter's API.

Practice 6. Dealing with bad formatted data

Practice to stress the importance of following best practices in data formatting.

Practice 7a. Locating HTML elements with bs4

Practice 7 is divided into 3 small steps to avoid confusion. So far, we have perfectly structured and formatted data, using different APIs. The goal now is to download data directly from a website, structure it and store it into a local file.

Practice 7b. Extracting data with bs4

Idem

Practice 7c. Storing data into a local file with bs4

Idem

Practice 8. Table parsing with requests and bs4

The goal is to make the students repeat the workflow seen in Practice #7, with data that comes inside an HTML <table> tag

Practice 9. Spider to grab every university profile inside of Wikipedia

Using already explained libraries and resources, the student is encouraged to build a Spider that recursively connects and retrieves all university profiles inside of Wikipedia.

Practice 10. Crawler to retrieve all useful data from the universities in Wikipedia

The student is asked to follow on with the code developed during the previous Practice #9 and concatenate a Crawler into it to collect the available data in the URLs that were already collected.

Practice 11. Web scraping legality

Theory regarding web scraping legal issues, built into a quiz to catch (engineering) students attention into important legal issues they might face.

Practice 12. Electric bill from the CNMC

Gathering public data regarding the electricity bill of any postal code in Spain from a public website, that forces interaction with HTML forms, and must be done with Selenium and Chromedriver

Practice 13. Spanish company house

So far, all practices ended downloading data. In this practice the goal is to download the PDFs listed in the Spanish Company House to make the student used to the Spanish definition of public transparency = a PDF.

Practice 14. Structuring unstructured data

The goal in this last practice is to parse the data contained into the PDF files downloaded in the previous practice #13, to be able to make any use out of them.

---

[1]

METODOLOGÍA DOCENTE

Aspectos metodológicos generales de la asignatura

To ensure useful and practical learning, theoretical classes will be combined with master classes that reflect the reality of the market. Real case studies with also be studied from business and technical perspectives, some of which will be used in practical sessions.

Metodología Presencial: Actividades

Lectures: The lecturer will introduce the fundamental concepts of each unit, along with some practical recommendations, and will go through worked examples to support the explanation. Active participation will be encouraged by raising open questions to foster discussion and by proposing quizzes and short application exercises to be solved in class.

Practical sessions: Under the instructor’s supervision, students will apply the concepts learned in the lectures to real cases, in order to face and solve implementation problems that typically arise.

Tutoring for groups or individual students will be organized upon request.

Metodología No presencial: Actividades

Personal study of the course material and resolution of the proposed exercises.

Practical session preparation to make the most of in-class time.

Practical results analysis and report writing.

RESUMEN HORAS DE TRABAJO DEL ALUMNO

STUDENT WORK-TIME SUMMARY	
IN-CLASS HOURS	
Lectures	Practical sessions
13	13

OUT-OF-CLASS HOURS			
Self-study	**Practice preparation**	**Report writing**	**Homework assignments**
12	0	10	42
		ECTS credits:	**3 (90 hours)**

EVALUACIÓN Y CRITERIOS DE CALIFICACIÓN

Assessment activities	Grading criteria	Weight
Final exam	▪ Understanding of the theoretical concepts. ▪ Application of these concepts to problem-solving. ▪ Critical analysis of numerical exercises' results.	30%
Practical assignments	▪ Application of theoretical concepts to real problem-solving. ▪ Ability to understand results in real environment. ▪ Written communication skills.	70%

Calificaciones

Grading

Regular assessment

- **Theory** will account for 30% and will be graded with an exam at the end of the subject.
- **Lab** (practical assignments) will account for the remaining 70%

To pass the course, the weighted average mark must be greater or equal to 5 out of 10 points, the mark of the final exam must be greater or equal to 5 out of 10 points.

Retake

- **Theory** will account for 30% and will be graded with an exam at the end of the subject.
- **Lab** (practical assignments) will account for the remaining 70%

To pass the course, the weighted average mark must be greater or equal to 5 out of 10 points, the mark of the final exam must be greater or equal to 5 out of 10 points.

Course rules

- Class attendance is mandatory according to Article 93 of the General Regulations (Reglamento General) of Comillas

Pontifical University and Article 6 of the Academic Rules (Normas Académicas) of the ICAI School of Engineering. Not complying with this requirement may have the following consequences:

  - Students who fail to attend more than 15% of the lectures may be denied the right to take the final exam during the regular assessment period.
  - Regarding practice, absence to more than 15% of the sessions can result in losing the right to take the final exam of the regular assessment period and the retake. Missed sessions must be made up for credit.

- Students who commit an irregularity in any graded activity will receive a mark of zero in the activity and disciplinary procedure will follow (cf. Article 168 of the General Regulations (Reglamento General) of Comillas Pontifical University).

PLAN DE TRABAJO Y CRONOGRAMA

Actividades

In and out-of-class activities	Date/Periodicity	Deadline
Final exam	After the lecture period	–
Practice sessions	During and after each lesson	–
Review and self-study of the concepts covered in the lectures	After each lesson	–
Practice preparation	Before every lab session	–
Practice report writing	–	Exam date

	In-class activities				Out-of-class activities				Learning outcomes
Week	Time [h]	Lecture	Laboratory	Assessment	Time [h]	Self-study	Practice preparation and report writing	Other activities	Code
1	2	Lecture 0 + Lecture I a)	Practice I		2	Review and self-study (2h)			
	2	Lecture I b)		Practice II	3.5		Practice III (7h)		

2	2	Lecture 2 a)	Practice IV		3	Review and self-study (2h)			
	2	Lecture 2 b)		Practice V	4.5		Practice V (7h)		
3	2	Lecture 3 a)	Practice VI		2.5	Review and self-study (2h)			
	2	Lecture 3 b)		Practice VII	4		Practice VIII (7h)		
4	2	Lecture 4 a)		Practice IX	6		Practice IX (7h)		
	2	Lecture 4 b)		Practice X	2		Practice X (7h)		
5	2	Lecture 5 a)	Practice XI		2	Review and self-study (2h)			
	2	Lecture 5 b)	Practice XII		2.5	Review and self-study (2h)			
6	2	Lecture 6 a)	Practice XIII		4		Practice XIII (7h)		
	2	Lecture 6 b)	Practice XIV		4.5	Review and self-study (2h)			
7	2	Q&A			2.5				
				Final exam[1]	10	Final exam preparation (10h)			

[1] The final exam will be held on the first week of December.

BIBLIOGRAFÍA Y RECURSOS

Bibliografía Básica

- Slides prepared by the lecturer
- Official documentation of the open source libraries: requests, bs4, selenium and chromedriver.
- Web scraping with Python, Ryan Mitchel, O'reilly. ISBN: 1491985577

Bibliografía Complementaria

Python Crash Course, Eric Matthes. ISBN: 1593279280