ANX-PR/CL/001-01

LEARNING GUIDE

SUBJECT

103000901 - Information Retrieval, Extraction And Integration

DEGREE PROGRAMME

10BA - Master Universitario En Ciencia De Datos

ACADEMIC YEAR & SEMESTER

2021/22 - Semester 2

Index

Learning guide

1. Description

1.1. Subject details

Name of the subject	103000901 - Information Retrieval, Extraction And Integration
No of credits	4.5 ECTS
Type	Optional
Academic year ot the programme	First year
Semester of tuition	Semester 2
Tuition period	February-June
Tuition languages	English
Degree programme	10BA - Master Universitario en Ciencia de Datos
Centre	10 - Escuela Tecnica Superior De Ingenieros Informaticos
Academic year	2021-22

2. Faculty

2.1. Faculty members with subject teaching role

Name and surname	Office/Room	Email	Tutoring hours *
Miguel Garcia Remesal (Subject coordinator)	2206	miguel.garcia.remesal@upm.es	Tu - 11:00 - 14:00 Th - 11:00 - 14:00
David Perez Del Rey	2104	david.perez.rey@upm.es	M - 14:00 - 16:00 W - 14:00 - 16:00 F - 11:00 - 13:00
M. Carmen Suarez De Figueroa Baonza	3205	mdelcarmen.suarezdefigueroa@upm.es	Tu - 14:00 - 16:00 W - 11:00 - 13:00 Th - 14:00 - 16:00

Victor Manuel Maojo Garcia	2102	victormanuel.maojo@upm.es	Tu - 12:30 - 15:30 W - 12:30 - 15:30
Raul Alonso Calvo		raul.alonso@upm.es	Sin horario.

* The tutoring schedule is indicative and subject to possible changes. Please check tutoring times with the faculty member in charge.

3. Skills and learning outcomes *

3.1. Skills to be learned

CECD01 - Conocer los procesos de captura, extracción, manipulación y conversión de datos en diferentes entornos.

CG06 - Especificación y realización de tareas informáticas complejas, poco definidas o no familiares

CG07 - Aplicación de los últimos o más novedosos métodos para resolver problemas que, posiblemente, involucren a otras disciplinas

CG09 - Integración del conocimiento de distintos campos de estudio

CG14 - Capacidad de trabajar y comunicarse también en contextos internacionales

3.2. Learning outcomes

RA32 - Understand and design information extraction systems

RA21 - Conocer cómo se aplican las técnicas de computación científica en algún campo específico de ciencia o ingeniería

RA33 - understand and apply information retrieval systems

RA34 - Apply AI techniques in real world data scenarios

* The Learning Guides should reflect the Skills and Learning Outcomes in the same way as indicated in the Degree Verification Memory. For this reason, they have not been translated into English and appear in Spanish.

4. Brief description of the subject and syllabus

4.1. Brief description of the subject

The amount of available data in most scientific areas has grown dramatically during the last few years. However, this increment did not have a parallel impact on the knowledge available for decision-making. There is a need for automated models to manage such data, considering that human beings will never directly use most of it. The course Information Retrieval, Extraction, and Integration focuses on the necessary methods and tools to extract information and models to efficiently retrieve data for further integration. These are critical tasks to provide relevant information for decision making, whose complexity increases with the amount of data available. As application areas, we focus mainly on biomedicine due to the complexity and the specific requirements.

4.2. Syllabus

1. Basic concepts
    1.1. Introduction
    1.2. Data, information and knowledge
    1.3. Data types
2. Handling textual data / Information retrieval
    2.1. Modern information retrieval
    2.2. Information extraction and Text Mining
3. Handling non-textual data
    3.1. Introduction and basic descriptors
    3.2. Content-based information retrieval
4. Data Integration
    4.1. Introduction
    4.2. Bias challenges
    4.3. Fairness challenges
5. Search engines
    5.1. Web search engines
    5.2. Machine learning-based ranking

6. Applications in biomedicine

6.1. Biomedical information systems

6.2. Biomedical vocabularies

6.3. Standars for clinical interoperability

6.4. Systems for retrieving scientific literature

5. Schedule

5.1. Subject schedule*

Week	Face-to-face classroom activities	Face-to-face laboratory activities	Distant / On-line	Assessment activities
1	**Presentación del curso** Duration: 01:00 **Desarrollo del tema 1.1** Duration: 01:00 **Desarrollo del Tema 1.2** Duration: 01:00 **Desarrollo del Tema 1.3** Duration: 01:00			**Estudio autónomo** Continuous assessment Not Presential Duration: 04:00 **Estudio autónomo** Continuous assessment Not Presential Duration: 04:00
2	**Desarrollo del Tema 1.3** Duration: 01:00 **Desarrollo del Tema 2.1** Duration: 02:00 **Desarrollo del Tema 2.2** Duration: 01:00			**Estudio autónomo** Continuous assessment Not Presential Duration: 04:00 **Estudio autónomo** Continuous assessment Not Presential Duration: 05:00
3	**Desarrollo del Tema 2.2** Duration: 01:00 **Desarrollo del Tema 2.3** Duration: 01:00 **Desarrollo del Tema 2.3** Duration: 02:00			**Entrega trabajo individual** Continuous assessment and final examination Not Presential Duration: 06:00 **Estudio autónomo** Continuous assessment Not Presential Duration: 05:00
4	**Desarrollo del Tema 2.4** Duration: 02:00			**Estudio autónomo** Continuous assessment Not Presential Duration: 04:00 **Estudio autónomo** Continuous assessment Not Presential Duration: 04:00

5	**Desarrollo del Tema 2.5** Duration: 01:00 **Trabajo en grupo supervisado** Duration: 01:00 **Desarrollo del Tema 2.5** Duration: 02:00			**Entrega trabajo en grupo** Continuous assessment and final examination Not Presential Duration: 06:00
6	**Desarrollo del Tema 3.1** Duration: 04:00			**Estudio autónomo** Continuous assessment Not Presential Duration: 03:00 **Estudio autónomo** Continuous assessment Not Presential Duration: 04:00
7	**Desarrollo del Tema 3.1** Duration: 01:00 **Desarrollo del Tema 3.2** Duration: 02:00 **Desarrollo del Tema 3.3** Duration: 01:00			**Estudio autónomo** Continuous assessment Not Presential Duration: 04:00 **Entrega de trabajo en grupo** Continuous assessment and final examination Presential Duration: 06:00
8	**Desarrollo del Tema 3.3** Duration: 02:00 **Desarrollo del Tema 3.4** Duration: 02:00			**Estudio autónomo** Continuous assessment Not Presential Duration: 02:00 **Estudio autónomo** Continuous assessment Not Presential Duration: 04:00
9				
10				
11				
12				
13				
14				
15				
16				

17				**Entrega de trabajo en grupo** Continuous assessment and final examination Presential Duration: 06:00

Depending on the programme study plan, total values will be calculated according to the ECTS credit unit as 26/27 hours of student face-to-face contact and independent study time.

* The schedule is based on an a priori planning of the subject; it might be modified during the academic year, especially considering the COVID19 evolution.

6. Activities and assessment criteria

6.1. Assessment activities

6.1.1. Continuous assessment

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
1	Estudio autónomo		No Presential	04:00	%	3 / 10	
1	Estudio autónomo		No Presential	04:00	%	3 / 10	
2	Estudio autónomo		No Presential	04:00	%	3 / 10	
2	Estudio autónomo		No Presential	05:00	%	3 / 10	
3	Entrega trabajo individual		No Presential	06:00	25%	3 / 10	CG07
3	Estudio autónomo		No Presential	05:00	%	3 / 10	
4	Estudio autónomo		No Presential	04:00	%	3 / 10	
4	Estudio autónomo		No Presential	04:00	%	3 / 10	
5	Entrega trabajo en grupo		No Presential	06:00	25%	3 / 10	CG07 CECD01
6	Estudio autónomo		No Presential	03:00	%	3 / 10	
6	Estudio autónomo		No Presential	04:00	%	3 / 10	
7	Estudio autónomo		No Presential	04:00	%	3 / 10	
7	Entrega de trabajo en grupo		Face-to-face	06:00	25%	3 / 10	CG14 CG09
8	Estudio autónomo		No Presential	02:00	%	3 / 10	
8	Estudio autónomo		No Presential	04:00	%	3 / 10	
17	Entrega de trabajo en grupo		Face-to-face	06:00	25%	3 / 10	CG07 CG06

6.1.2. Final examination

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
3	Entrega trabajo individual		No Presential	06:00	25%	3 / 10	CG07
5	Entrega trabajo en grupo		No Presential	06:00	25%	3 / 10	CG07 CECD01
7	Entrega de trabajo en grupo		Face-to-face	06:00	25%	3 / 10	CG14 CG09
17	Entrega de trabajo en grupo		Face-to-face	06:00	25%	3 / 10	CG07 CG06

6.1.3. Referred (re-sit) examination

No se ha definido la evaluación extraordinaria.

6.2. Assessment criteria

Ordinary call

The student can choose the evaluation system in the ordinary call, either continuous evaluation or final test. Students wishing to take the final test modality **MUST** notify the course's coordinator (mgremesal@fi.upm.es) **WITHIN THE FIRST 15 DAYS** from the beginning of the teaching activity.

Continuous evaluation modality

The subject's final grade will be calculated from the qualifications of an individual assignment and 3 group assignments.

Individual work. After presenting the "Handling textual data / Information Retrieval" unit, students must do an assignment that may be presented in class if required by the instructor. The qualification of this work will be 25% of the final grade.

Group work on the application of data extraction from images. An implementation, a brief report, and a presentation in class should be made. The qualification of this work will be 25% of the final grade.

Group Work on search engines. The qualification of this work will be 25% of the final grade.

Group Work on Data Integration. The qualification of this work will be 25% of the final grade.

Final test modality

The student must pass a written exam covering the entire syllabus. To pass the exam, the student is expected to obtain a grade equal to or higher than 5 (out of 10) points.

Extraordinary call

In the extraordinary call, it is only possible to take the final test modality.

7. Teaching resources

7.1. Teaching resources for the subject

Name	Type	Notes
Modern Information Retrieval	Bibliography	Baeza-Yates, Ricardo, and Berthier Ribeiro-Neto. Modern information retrieval. New York: ACM press, 1999.
The data warehouse toolkit	Bibliography	Kimball, Ralph, and Margy Ross. The data warehouse toolkit: the complete guide to dimensional modeling. John Wiley & Sons, 2011.
Introduction to Information Retrieval	Bibliography	Manning CD, Raghavan P, Schütze H. Introduction to Information Retrieval. Cambridge University Press. 2008
Managing Gigabytes	Bibliography	Witten IH, Moffat A, Bell TC. Managing Gigabytes: Compressing and Indexing Documents and Images, 2nd Edition. Morgan Kaufmann. 1999.

Natural Language Processing with Python	Bibliography	7. Bird S, Klein E, Loper E. Natural Language Processing with Python. O'Reilly 2009.In successive academic years the individual work prepared by E1 students will be also available for other students? cohorts.