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

2022/23 - 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	2022-23

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
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
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

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. IR systems evaluation
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	Classroom activities	Laboratory activities	Distant / On-line	Assessment activities
1	**Course presentation** Duration: 01:00 **Topic 1** Duration: 01:00 **Topic 2** Duration: 02:00			
2	**Topic 2** Duration: 04:00			**Group assignment submission (topic 2) and class presentation (under request of the instructor)** Continuous assessment and final examination Presential Duration: 03:00
3	**Topic 3** Duration: 04:00			
4	**Topic 3** Duration: 02:00 **Topic 4** Duration: 02:00			**Group assignment submission (topic 3) and class presentation** Continuous assessment and final examination Presential Duration: 03:00
5	**Topic 4** Duration: 04:00			**Group assignment submission (topic 4) and class presentation (under request of the instructor)** Continuous assessment and final examination Presential Duration: 03:00
6	**Topic 5** Duration: 04:00			
7	**Topic 5** Duration: 02:00 **Topic 6** Duration: 02:00			**Group assignment submission (topic 5) and class presentation (under request of the instructor)** Continuous assessment and final examination Presential Duration: 03:00

8	**Topic 6** Duration: 04:00			
9				
10				
11				
12				
13				
14				
15				
16				
17				

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. Assessment

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
2	Group assignment submission (topic 2) and class presentation (under request of the instructor)		Face-to-face	03:00	25%	5 / 10	CECD01 CG14
4	Group assignment submission (topic 3) and class presentation		Face-to-face	03:00	25%	5 / 10	CG07 CG14
5	Group assignment submission (topic 4) and class presentation (under request of the instructor)		Face-to-face	03:00	25%	5 / 10	CG07 CG14
7	Group assignment submission (topic 5) and class presentation (under request of the instructor)		Face-to-face	03:00	25%	5 / 10	CG09 CG14

6.1.2. Global examination

Week	Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
2	Group assignment submission (topic 2) and class presentation (under request of the instructor)		Face-to-face	03:00	25%	5 / 10	CECD01 CG14
4	Group assignment submission (topic 3) and class presentation		Face-to-face	03:00	25%	5 / 10	CG07 CG14
5	Group assignment submission (topic 4) and class presentation (under request of the instructor)		Face-to-face	03:00	25%	5 / 10	CG07 CG14
7	Group assignment submission (topic 5) and class presentation (under request of the instructor)		Face-to-face	03:00	25%	5 / 10	CG09 CG14

6.1.3. Referred (re-sit) examination

Description	Modality	Type	Duration	Weight	Minimum grade	Evaluated skills
Final exam		Face-to-face	01:00	100%	5 / 10	CECD01 CG09 CG06 CG07 CG14

6.2. Assessment criteria

Progressive evaluation (during the class period)

The final grade will be calculated from the marks of four mandatory group practical assignments on topics 2, 3, 4 and 5. Each practical assignment weights equally (25%) on the final grade, so it will be calculated as the arithmetic mean of all the assignments' grades. In addition, the faculty responsible for each assignment may request the groups to carry out a GROUP presentation in the classroom about the work done in such assignment. This presentation is mandatory for one of the assignments (topic 3) and at the discretion of the instructor responsible of the assignment for the rest of topics. The mark for the presentation will be integrated into the grade of the assignment and thus it may differ between members of the same group.

To pass the course, the individual grades for ALL the proposed practical assignments must be equal to or greater than 5 points. However, it will also be possible to pass the course if at most one of the assignments has been graded with a mark lower than 5 points and equal to or greater than 3 points. In the latter case, the final grade will be calculated by subtracting 1 point from the arithmetic mean calculated with a floor of 5 points. Therefore, if the resulting grade after the deduction is less than 5 points, then the student's final grade would be 5 points (pass). In case of having obtained a grade lower than 5 points in two or more assignments, then the grade will be calculated as the minimum between 4 points and the arithmetic mean of the grades for all assignments (fail in both cases)

Any practical assignment can be submitted on the scheduled date even if any of the previous works have not been delivered. The failed or non delivered assignments can be resubmitted in the ordinary call, although note that it will no longer be possible to obtain the highest grade (see regulations for the ordinary call)

Ordinary call

If the student failed to pass the course by progressive evaluation, then he/she will be allowed to submitt ALL the assignments that he/she has not delivered during the progressive evaluation, as well as those in which he/she has obtained a grade lower than 5 points. It will not be possible to make a new submission of assignments passed during the progressive evaluation (grade equal to or greater than 5 points). The delivery deadline will be the one established in the official calendar of the degree for the global test of the ordinary call. The student will have to make an INDIVIDUAL presentation in the classroom for each of the works presented in this call if the professor responsible for the work considers it so. These presentations, if eventually required, will be carried out in classroom on the date scheduled in the official calendar of the degree for the global test of the ordinary call. As in the progressive evaluation, the grade of the presentation, if required, will be integrated into the final grade of the assignment. It will only be possible to pass the course if the grade for ALL the assignments delivered in this call, together with those passed during the progressive evaluation, is equal to or greater than 5 points. The final grade will be calculated as the arithmetic mean of the grades of all the assignments, substracting 0.5 points from this arithmetic mean for each assignment delivered in this call with a floor of 5 points. In case of having obtained a grade lower than 5 points in one or more assignments, then the grade will be calculated as the minimum between 4 points and the arithmetic mean of the grades of all practices (fail in both cases)

Extraordinary call

In the extraordinary call there will be a final written exam that will cover all the contents of the course. To pass the course it will be necessary to obtain a grade equal to or greater than 5 points. The final grade is the exam mark, graded out of 10 points.

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.

8. Other information

8.1. Other information about the subject

The course will be imparted intensively during half a semester (8 weeks)