ANX-PR/CL/001-01
LEARNING GUIDE

SUBJECT

103000829 - Information Retrieval Extraction And Integration

DEGREE PROGRAMME

10AX - Master Universitario Innovaci?n Digital Ciencia de Datos Itinerario Health

ACADEMIC YEAR & SEMESTER

2020/21 - Semester 2

Index

Learning guide

1. Description

1.1. Subject details

Name of the subject	103000829 - 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	10AX - Master Universitario Innovación Digital Ciencia de Datos Itinerario Health
Centre	10 - Escuela Tecnica Superior de Ingenieros Informaticos
Academic year	2020-21

2. Faculty

2.1. Faculty members with subject teaching role

Name and surname	Office/Room	Email	Tutoring hours *
Miguel Garcia Remesal	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 (Subject coordinator)	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

CB07 - Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio

CB10 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

CE-HMDA06 - Capacidad para extraer, integrar y consultar datos heterogéneos en escenarios clínicos

CE-HMDA07 - Capacidad para diseñar y gestionar proyectos de salud y datos médicos

CG01 - Que los estudiantes sean 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.

CG03 - La capacidad de usar la lengua inglesa de manera competente, es decir, con capacitación para tareas complejas de trabajo y estudio.

CG06 - Capacidad para gestionar la información.

3.2. Learning outcomes

RA33 - Understand and design information extraction systems

RA3 - Conocer cómo se aplican las técnicas de computación científica en algún campo específico de ciencia o ingeniería

RA34 - Understand and apply information retrieval systems

* 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 any area has grown dramatically during the las years. However, this increment did not have a proportional impact in the knowledge available for decision making. There is a need of automatic models to manage the data, taking into account that the majority of the data will never be used by a human being. The course Information Retrieval, Extraction and Integration is focused on the necessary tasks to extract information, models to efficiently retrieve data for further integration. These are critical tasks to provide relevant information for decision making, which complexity increases with the amount of data available. As application areas, we focus on biomedicine, due to the complexity and to the specific requriments.

4.2. Syllabus

1. Basic Concepts
    1.1. Introduction
    1.2. Data, Information and Knowledge
    1.3. Data types
2. Extraction and Information Retrieval
    2.1. Information Extraction
    2.2. Information Retrieval Models
    2.3. Natural Language Processing
    2.4. Web Search Engines
    2.5. Non-textual Data
3. Data Integration
    3.1. Integration Architectures
    3.2. Semantic Interoperability
    3.3. Data Provenance
4. Applications in Biomedicine
    4.1. Sistemas de información biomédica
    4.2. Estándares de interoperabilidad clínica
    4.3. Vocabularios biomédicos
    4.4. Sistemas de recuperación de literatura científica

5. Schedule

5.1. Subject schedule*

Week	Face-to-face classroom activities	Face-to-face laboratory activities	Distant / On-line	Assessment activities
1	**Presentación de la asignatura** 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: 01:00 **Desarrollo del Tema 2.1** Duration: 01:00 **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 2.3** Duration: 01:00 **Desarrollo del Tema 2.4** Duration: 01:00			**Estudio autónomo** Continuous assessment Not Presential Duration: 04:00 **Estudio autónomo**

	Desarrollo del Tema 2.5 Duration: 01:00 **Trabajo en grupo supervisado** Duration: 01:00			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 3.1** Duration: 02:00			**Entrega trabajo en grupo** Continuous assessment and final examination Not Presential Duration: 06:00 **Estudio autónomo** Continuous assessment Not Presential Duration: 03:00
6	**Desarrollo del Tema 3.1** Duration: 02:00 **Desarrollo del Tema 3.2** Duration: 02:00			**Estudio autónomo** Continuous assessment Not Presential Duration: 04:00 **Estudio autónomo** Continuous assessment Not Presential Duration: 04:00
7	**Desarrollo del Tema 3.2** Duration: 01:00 **Desarrollo del Tema 3.3** Duration: 01:00 **Desarrollo del Tema 3.3** Duration: 02:00			**Entrega de trabajo en grupo** Continuous assessment and final examination Presential Duration: 06:00 **Estudio autónomo** Continuous assessment Not Presential Duration: 02:00
8	**Desarrollo del Tema 3.4** Duration: 04:00			**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	CB07 CE-HMDA07
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	CE-HMDA06 CB10
5	Estudio autónomo		No Presential	03:00	%	3 / 10	
6	Estudio autónomo		No Presential	04:00	%	3 / 10	
6	Estudio autónomo		No Presential	04:00	%	3 / 10	
7	Entrega de trabajo en grupo		Face-to-face	06:00	25%	3 / 10	CE-HMDA06 CG06 CG01
7	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	CG03 CB10

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	CB07 CE-HMDA07
5	Entrega trabajo en grupo		No Presential	06:00	25%	3 / 10	CE-HMDA06 CB10

7	Entrega de trabajo en grupo		Face-to-face	06:00	25%	3 / 10	CE-HMDA06 CG06 CG01
17	Entrega de trabajo en grupo		Face-to-face	06:00	25%	3 / 10	CG03 CB10

6.1.3. Referred (re-sit) examination

No se ha definido la evaluación extraordinaria.

6.2. Assessment criteria

Ordinary evaluation

The final grade of the subject will be calculated from the qualifications of an individual work and 3 group work.

Individual work. After the presentation of the first unit of text mining, students must do a assignment. 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 Semantic interoperability. The qualification of this work will be 25% of the final grade.

Modality of extraordinary evaluation

In the ordinary call, the choice between the system of continuous evaluation or the system of evaluation by final test corresponds to the student. Anyone wishing to follow the evaluation system by means of a final test only, must MUST notify it DURING THE FIRST 15 DAYS from the beginning of the teaching activity of the subject, by writing to the Prof. Coordinator of the subject to be delivered within the established deadline. The qualification in case of final evaluation will be through exam.

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

Course will be given in 8 weeks. If the COVID19 crisis requires it, it will be fully online.