Teaching Guide				
Identifying Data				2022/23
Subject (*)	Natural Language Processing and Text Mining		Code	614G02043
Study programme	Grao en Ciencia e Enxeñaría de Datos			
Descriptors				
Cycle	Period	Year	Type	Credits
Graduate	2nd four-month period	Fourth	Optional	6
Language	Spanish			
Teaching method	Face-to-face			
Prerequisites				
Department	Ciencias da Computación e Tecnoloxías da Información			
Coordinador	Vilares Calvo, David	E-mail	david.vilares@udc.es	
Lecturers	Gómez Rodríguez, Carlos Vilares Calvo, David	E-mail	carlos.gomez@udc.es david.vilares@udc.es	
Web	campusvirtual.udc.es			
General description	Natural language processing (NLP) is the area of artificial intelligence that deals with the study and development of computational models that are capable of processing and understanding the particularities of natural language as efficiently as humans do. In this course, students will be introduced to basic fundamentals and machine learning techniques associated with NLP, which are used as a starting point for the development of numerous language technologies and automatic text mining. Students will become familiar with algorithms and techniques for representing as trees and graphs the latent information present in written texts, with techniques for representing words in a way that efficiently captures their meaning, with the implementation of models using machine learning techniques for their application to various NLP-related problems that are useful for non-specialized users, as well as with existing techniques for language technologies so they can be applied to many languages, even those for which there is a limited amount of available resources (annotated or not).			

Study programme competences	
Code	Study programme competences
A28	CE28 - Comprensión e dominio dos fundamentos e técnicas para o procesado de datos escritos, tanto en linguaxe formal como en linguaxe natural.
B2	CB2 - Que os estudantes saiban aplicar os seus coñecementos ao seu traballo ou vocación dunha forma profesional e posúan as competencias que adoitan demostrarse por medio da elaboración e defensa de argumentos e a resolución de problemas dentro da súa área de estudo
B3	CB3 - Que os estudantes teñan a capacidade de reunir e interpretar datos relevantes (normalmente dentro da súa área de estudo) para emitir xuízos que inclúan unha reflexión sobre temas relevantes de índole social, científica ou ética
B4	CB4 - Que os estudantes poidan transmitir información, ideas, problemas e solucións a un público tanto especializado como non especializado
B7	CG2 - Elaborar adecuadamente e con certa orixinalidade composicións escritas ou argumentos motivados, redactar plans, proxectos de traballo, artigos científicos e formular hipóteses razoables.
B8	CG3 - Ser capaz de manter e estender formulacións teóricas fundadas para permitir a introdución e explotación de tecnoloxías novas e avanzadas no campo.
B9	CG4 - Capacidade para abordar con éxito todas as etapas dun proxecto de datos: exploración previa dos datos, preprocesado, análise, visualización e comunicación de resultados.
B10	CG5 - Ser capaz de traballar en equipo, especialmente de carácter multidisciplinar, e ser hábiles na xestión do tempo, persoas e toma de decisións.

Learning outcomes

Learning outcomes	Study programme competences		
To know, understand and analyze natural language processing techniques for processing and disambiguation at syntactic and semantic levels.	A28	B2 B3 B4 B7 B8 B9 B10	
To know how to use the techniques and methods of natural language processing to solve real text mining problems.	A28	B2 B3 B4 B7 B8 B9 B10	
To know and understand the problems posed by multilingualism in data sources and techniques to solve them.	A28	B2 B8 B9 B10	
To know and analyze emerging computing paradigms with the potential to improve parallelism in text mining.	A28	B2 B4 B7 B8	

Contents	
Topic	Sub-topic
Constituent parsing for text mining	Syntax of constituents Statistical constituent analysis with dynamic programming Analysis of shift-reduce constituents with neural networks Analysis of discontinuous constituents Sequence-by-sequence constituent analysis
Dependency parsing for text mining	Dependency Syntax Annotation criteria and universal dependencies Dependency analysis based on transitions Analysis of dependencies based on graphs Non-projectivity
Semantics	Analysis of semantic dependencies Dense vectors using SVD Dense vectors using word prediction: skip-gram and CBOW Properties of dense vectors Brown clustering
Computing with word senses	Word senses Relations between senses Databases of lexical relationships Disambiguation of the meaning of words
Practical applications of text mining	-

Multilingual language processing	Processing of morphologically-rich languages Non-segmented language processing Language processing with few resources Translingual processing
Emerging technologies	-

Planning				
Methodologies / tests	Competencies	Ordinary class hours	Student?s personal work hours	Total hours
Guest lecture / keynote speech	A28 B3 B8 B9	21	21	42
Laboratory practice	A28 B2 B4 B7 B9 B10	10	30	40
Problem solving	A28 B2 B3 B8	8	30	38
Objective test	A28 B2 B3 B8	2	10	12
Supervised projects	A28 B3 B4 B7 B10	1	16	17
Personalized attention		1	0	1
(*)The information in the planning table is for guidance only and does not take into account the heterogeneity of the students.				

Methodologies	
Methodologies	Description
Guest lecture / keynote speech	Theoretical classes, in which the content of each topic is exposed. The student will have copies of the slides in advance and the teacher will promote an active attitude, asking questions that allow clarifying specific aspects and leaving questions open for the student's reflection.
Laboratory practice	Practical classes with the use of a computer, which allow the student to familiarize himself/herself from a practical point of view with the issues exposed in the theoretical classes.
Problem solving	Problem-based learning, seminars, case studies and projects.
Objective test	Mastery of theoretical and operational knowledge of the subject will be assessed.
Supervised projects	Work in which students will consult sources of information to become familiar with research aspects of the field

Personalized attention	
Methodologies	Description
Objective test Guest lecture / keynote speech Laboratory practice Supervised projects Problem solving	The development of the master classes, as well as of the problem solving classes and the practical laboratories, will be carried out according to the progress of the students in the comprehension and assimilation of the contents taught. The general progress of the class will be combined with a specific attention to those students who present greater difficulties in the task of learning and with an additional support to those who present greater fluency and wish to broaden their knowledge. In supervised projects, personalized attention will be provided to students to guide them in their autonomous work. With regard to individual tutorials, given their personalized nature, they should not be devoted to extend the contents with new concepts, but to clarify the concepts already exposed. The teacher will use them as an interaction that will allow them to draw conclusions regarding the degree of assimilation of the subject by the students.

Assessment			
Methodologies	Competencies	Description	**Qualification**
Objective test	A28 B2 B3 B8	Compulsory performance. Mastery of theoretical and operational knowledge of the subject will be assessed.	50
Laboratory practice	A28 B2 B4 B7 B9 B10	The deliveries of the practices must be made within the period established in the virtual campus and must follow the specifications indicated in the statement both for their submission and their defense.	40

Supervised projects	A28 B3 B4 B7 B10	The students abilities to understand and assimilate research work will be evaluated.	10

Assessment comments
Students must achieve at least 40% of the maximum mark of the theory and practice parts, and in any case the sum of three parts must exceed 5 to pass the subject. If any of the above requirements is not met, the grade of the call will be established according to the lowest grade obtained.In case of not reaching the minimum in theory or practice, the student will have a second opportunity in which they will only be required to deliver said part. The tutored works are considered as continuous evaluation and will not be delivered in the second opportunity.Grades will not be saved between academic courses.The deliveries of the practices must be made within the period established in the virtual campus and must follow the specifications indicated in the statement both for their submission and their defense.Whoever attends the objective test in the official evaluation period will have the status of "Presented".In the case of fraudulent completion of exercises or tests, the Regulations for evaluating the academic performance of students and reviewing qualifications will be applied. In application of the corresponding regulations on plagiarism, the total or partial copy of some practice or theory exercise will suppose the suspense in the two opportunities of the course, with the qualification of 0.0 in both cases.

Sources of information	
Basic	- Manning, C., & Schutze, H. (1999). Foundations of statistical natural language processing. MIT Press - Goldberg, Y. (2017). Neural network methods for natural language processing. Synthesis lectures on human language technologies. Morgan Claypool - Jacob Eisenstein (2019). Introduction to Natural Language Processing. MIT Press - Jurafsky, D. & Martin, J. H. (2022). Speech and Language Processing (3rd ed. draft). Disponible en: https://web.stanford.edu/~jurafsky/slp3/
Complementary	- Chollet, F. (2018). Keras: The python deep learning library. Astrophysics Source Code Library - Stuart Russell, Peter Norvig (2020). Artificial Intelligence: A Modern Approach, 4th Edition. Pearson - Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze (2008). Introduction to Information Retrieval. Cambridge University Press, Cambridge - Kübler, S., McDonald, R., & Nivre, J. (2009). Dependency Parsing. Synthesis lectures on human language technologies. Morgan Claypool

Recommendations
Subjects that it is recommended to have taken before
Written Language Processing/614G02029 Machine Learning III/614G02026 Machine Learning I/614G02019 Machine Learning II/614G02021
Subjects that are recommended to be taken simultaneously
Subjects that continue the syllabus
Other comments

(*)The teaching guide is the document in which the URV publishes the information about all its courses. It is a public document and cannot be modified. Only in exceptional cases can it be revised by the competent agent or duly revised so that it is in line with current legislation.