Teaching Guide				
Identifying Data				2023/24
Subject (*)	Language Modelling		Code	614544009
Study programme	Máster Universitario en Intelixencia Artificial			
Descriptors				
Cycle	Period	Year	Type	Credits
Official Master's Degree	2nd four-month period	First	Optional	3
Language	English			
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	Vilares Calvo, David	E-mail	david.vilares@udc.es	
Web	campusvirtual.udc.es			
General description	Provide theoretical knowledge that allows an in-depth study of linguistic models, such as language models and models of distributional semantics. Link language modeling and model types to different tasks within the area of language technologies and natural language processing. Evaluate different aspects of language models. Provide practical knowledge to train language models and use them in different natural language processing tasks.			

Study programme competences / results	
Code	Study programme competences / results
A2	CE01 - Understanding and command of techniques for lexical, syntactic and semantic processing of text in natural language
A3	CE02 - Understanding and command of fundamentals and techniques for processing linked documents, both structured and unstructured, and of the representation of their contents
A4	CE03 - Understanding and knowledge of the techniques for knowledge representation and processing for ontologies, graphs and RDF, together with their associated tools
B1	CG01 - Maintaining and extending theoretical foundations to allow the introduction and exploitation of new and advanced technologies in the field of AI
B3	CG03 - Searching and selecting that useful information required to solve complex problems, with a confident handling of bibliographical sources in the field
B4	CG04 - Suitably elaborating written essays or motivated arguments, including some point of originality, writing plans, work projects, scientific papers and formulating reasonable hypotheses in the field
B6	CB01 - Acquiring and understanding knowledge that provides a basis or opportunity to be original in the development and/or application of ideas, frequently in a research context
B7	CB02 - The students will be able to apply the acquired knowledge and to use their capacity of solving problems in new or poorly explored environments inside wider (or multidisciplinary) contexts related to their field of study
B10	CB05 - The students will acquire learning abilities to allow them to continue studying in way that will mostly be self-directed or autonomous
C2	CT02 - Command in understanding and expression, both in oral and written forms, of a foreign language
C3	CT03 - Use of the basic tools of Information and Communications Technology (ICT) required for the student's professional practice and learning along her life
C7	CT07 - Developing the ability to work in interdisciplinary or cross-disciplinary teams to provide proposal that contribute to a sustainable environmental, economic, political and social development
C8	CT08 - Appreciating the importance of research, innovation and technological development in the socioeconomic and cultural progress of society

Learning outcomes			
Learning outcomes	Study programme competences / results		
To know how to use the techniques and methods of natural language processing to solve real problems of analysis of texts in natural language.	AC1 AC3	BC1 BC3 BC4 BC7 BC10	CC2 CC3 CC7
To know, understand and analyze deep learning techniques applied to natural language processing.	AC1 AC2	BC1 BC3 BC6 BC7 BC10	CC2 CC3 CC7 CC8
To know how to use deep learning techniques and methods to solve practical problems in natural language processing.	AC1 AC2	BC1 BC3 BC4 BC6 BC7 BC10	CC2 CC3 CC7 CC8
To know and understand the environmental problems posed by the computational cost of deep learning techniques when applied to text analysis	AC1 AC2	BC1 BC6	CC2 CC8

Contents	
Topic	Sub-topic
Language models	N-gram based language models Neural based language models
Distributional semantics models	Linguistic hypothesis about distributional meaning Classic models of distributional semantics Neural models representing static meaning (word embeddings) Neural models representing dynamic-contextual meaning Compositional models
Sequence labeling	Use and fine-tuning of models for sequence labeling
Text-To-Text models	Uso e adaptación de modelos para o etiquetado secuencial

Planning				
Methodologies / tests	Competencies / Results	Teaching hours (in-person & virtual)	Student?s personal work hours	Total hours
Guest lecture / keynote speech	A2 A3 A4 B1 B3 B6 B7 B10 C2 C8	10	10	20
Laboratory practice	A2 A3 B3 B4 B6 B7 B10 C2 C3 C7 C8	5	17	22
Problem solving	A2 A3 B3 B4 B6 B7 B10 C2 C8	6	15	21
Multiple-choice questions	A2 A3 B1 B6 B7 B10 C2	0	1	1
Objective test	A2 A3 B1 B6 B7 B10 C2 C3	2	8	10
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.
Multiple-choice questions	Brief questionnaires to be filled after some theoretical sessions to help assimilate the content of the lecture.
Objective test	The mastery of the theoretical and operating knowledge of the subject will be evaluated.

Personalized attention	
Methodologies	**Description**
Laboratory practice Problem solving Objective test Guest lecture / keynote speech	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. 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 / Results**	**Description**	**Qualification**
Laboratory practice	A2 A3 B3 B4 B6 B7 B10 C2 C3 C7 C8	The deliveries of the practices must be made within the period established in the virtual campus and must follow the specifications indicated in the assigment both for their presentation and their defense.	50
Objective test	A2 A3 B1 B6 B7 B10 C2 C3	Compulsory. Mastery of theoretical and operational knowledge of the subject will be assessed.	45
Multiple-choice questions	A2 A3 B1 B6 B7 B10 C2	Small continuous assessment questionnaires that will be proposed at the end of some theoretical sessions and where you will be asked in a simple way about some of the concepts explained in that session. It will be notified in advance.	5

Assessment comments

Students must achieve a minimum of 40% of the maximum mark of the "Laboratory Practices" and "Objective Test" parts, and in any case the sum of the three parts must be greater than 5 to pass the subject. If any of the above requirements is not met, the grade for the course will be established according to the lowest grade obtained.
In case of not reaching the minimum score in the "Laboratory Practices" or "Objective Test" parts, the student will have a second opportunity in which only the delivery of the failed part will be required.
Grades will not be saved between academic years.
The delivery of the practicals must be done within the deadline established in the virtual campus and must follow the specifications indicated in the statement for both its presentation and defense.
The student who submits all the compulsory practicals or attends the objective test in the official evaluation period will be considered "Presented".
In the case of fraudulent completion of exercises or tests, the Regulations for the evaluation of the academic performance of students and review of grades will be applied. In application of the corresponding regulations on plagiarism, the total or partial copy of any practice or theory exercise will result in suspension on both occasions of the course, with a grade of 0.0 in both cases.

Sources of information	
Basic	Jurafsky, Daniel & James H. Martin (2021). ?N-gram Language Models.? Speech and Language Processing, Capítulo 3. https://web.stanford.edu/~jurafsky/slp3/Jurafsky, Daniel & James H. Martin (2021). ?Vector Semantics and Embeddings.? Speech and Language Processing, Capítulo 6. https://web.stanford.edu/~jurafsky/slp3/Jurafsky, Daniel & James H. Martin (2021). ?Neural Networks and Neural Language Models.? Speech and Language Processing, Capítulo 7. https://web.stanford.edu/~jurafsky/slp3/Jurafsky, Daniel & James H. Martin (2021). ?Sequence Labeling for Parts of Speech and Named Entities.? Speech and Language Processing, Capítulo 8. https://web.stanford.edu/~jurafsky/slp3/Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova (2018). ?BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171?4186, Minneapolis, Minnesota. Association for Computational Linguistics.Erk, Katrin (2012). "Vector space models of word meaning and phrase meaning: A survey." Language and Linguistics Compass 6.10: 635-653.
Complementary	Baroni, Marco, Raffaella Bernardi & Roberto Zamparelli (2014). ?Frege in space: A program for compositional distributional semantics.? Linguistic Issues in Language Technologies 9(6): 5-110.Baroni, Marco, Georgiana Dinu & Germán Kruszewski (2014). ?Don?t count, predict! A systematic comparison of context-counting vs. context-predicting semantic vectors.? In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) , pp. 238?247, Baltimore, Maryland. Association for Computational Linguistics.Church, Kenneth Ward, Zeyu Chen & Yanjun Ma (2021). ?Emerging trends: A gentle introduction to fine-tuning.? Natural Language Engineering, 27: 763?778.Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova (2018). ?BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 4171?4186, Minneapolis, Minnesota. Association for Computational Linguistics.Erk, Katrin (2012). "Vector space models of word meaning and phrase meaning: A survey." Language and Linguistics Compass 6.10: 635-653.Hirschberg, Julia & Christopher D. Manning (2015). "Advances in natural language processing." Science 349.6245: 261-266.Linzen, Tal (2016). "Issues in evaluating semantic spaces using word analogies." In Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP, pp. 13?18, Berlin, Germany. Association for Computational Linguistics.Mikolov, Tomas, Wen-tau Yih & Geoffrey Zweig (2013). "Linguistic Regularities in Continuous Space Word Representations." In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 746?751, Atlanta, Georgia. Association for Computational Linguistics.Taher Pilehvar, Mohammad & Jose Camacho-Collados (2021). Embeddings in Natural Language Processing: Theory and Advances in Vector Representations of Meaning. Morgan & Claypool (Synthesis Lectures on Human Language Technologies, volume 47).

Recommendations
Subjects that it is recommended to have taken before
Natural Language Understanding/614544008 Machine Learning I /614544012
Subjects that are recommended to be taken simultaneously
Deep Learning /614544013 Machine Learning II /614544014
Subjects that continue the syllabus
Text Mining/614544011 Web Intelligence and Semantic Technologies/614544010
Other comments
As stated in the different university teaching regulations, this subject incorporates the gender perspective. Student participation in class will be encouraged, and work will be done to identify and modify prejudices and sexist attitudes and influence the environment to modify them and promote values ??of respect and equality. Situations of discrimination based on gender must be detected and actions and measures to correct them will be proposed.