



Teaching Guide

Teaching Guide				
Identifying Data				2022/23
Subject (*)	Written Language Processing		Code	614G02029
Study programme	Grao en Ciencia e Enxeñaría de Datos			
Descriptors				
Cycle	Period	Year	Type	Credits
Graduate	2nd four-month period	Third	Obligatory	6
Language	SpanishGalician			
Teaching method	Face-to-face			
Prerequisites				
Department	Ciencias da Computación e Tecnoloxías da Información			
Coordinador	Dafonte Vazquez, Jose Carlos		E-mail	carlos.dafonte@udc.es
Lecturers	Alonso Pardo, Miguel angel		E-mail	miguel.alonso@udc.es
	Dafonte Vazquez, Jose Carlos			carlos.dafonte@udc.es
	Vilares Calvo, David			david.vilares@udc.es
Web	https://campusvirtual.udc.gal			
General description	A materia introduce nocións necesarias para procesar información, tanto en linguaxes formais empregadas para datos estruturados como en linguaxe humana. A materia inclúe tanto os fundamentos básicos do procesamento da linguaxe estruturada para o deseño de recoñecedores: análise léxica, sintáctica e semántica; como do procesamento da linguaxe natural: modelado da linguaxe, análise morfolóxica e sintáctica.			

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.
C1	CT1 - Utilizar as ferramentas básicas das tecnoloxías da información e as comunicacións (TIC) necesarias para o exercicio da súa profesión e para a aprendizaxe ao longo da súa vida.
C4	CT4 - Valorar a importancia que ten a investigación, a innovación e o desenvolvemento tecnolóxico no avance socioeconómico e cultural da sociedade.

Learning outcomes

Learning outcomes	Study programme competences
-------------------	-----------------------------



Coñecer, comprender e analizar a representación formal de diversos fenómenos léxicos, sintácticos e semánticos da linguaxe	A28	B3 B4 B7 B8 B10	C4
Coñecer, comprender e saber usar as tecnoloxías, marcos e bibliotecas para a construción de sistemas de procesamento da linguaxe	A28	B2 B3 B4 B7 B8 B9 B10	C1 C4
Deseñar, implementar e saber usar algoritmos e estruturas de datos para tratar e dar soporte aos diversos fenómenos característicos da linguaxe	A28	B2 B3 B4 B7 B8 B9 B10	C1 C4

Contents	
Topic	Sub-topic
Tema I. Fundamentos da linguaxe estruturada	1.1. Conceptos básicos: gramáticas e linguaxes 1.2. Fases e compoñentes dun procesador de linguaxe 1.3. Contornos de aplicación da Teoría de Linguaxes
Tema II. Análise léxica	2.1. Autómatas finitos, expresións regulares e gramáticas regulares 2.2. Conversión entre distintas representacións 2.3. Implementación de autómatas e expresións regulares
Tema III. Análise sintáctica	3.1. Linguaxes de contexto libre e gramáticas independentes do contexto 3.2. Recoñecedores descendentes recursivos e LL (k) 3.3. Recoñecedores LR (k)
Tema IV. Análise semántica e deseño de accións semánticas	4.1. Control de atributos 4.2. Tratamento de erros 4.3. Xeración dirixida pola sintaxe
Tema V. Introducción ao procesamento da linguaxe natural	5.1. Niveis de análise 5.2. Ambigüidade
Tema VI. Modelado da linguaxe	6.1. Modelos de linguaxe baseados en n-gramas 6.2. Avaliación do modelo de linguaxe 6.3. Técnicas de suavizado 6.4. Modelos de linguaxe baseados en redes neuronais
Tema VII. Análise morfolóxica da linguaxe natural	7.1. Fundamentos da análise morfolóxica 7.2. Análise morfolóxica estatística 7.3. Análise morfolóxica con redes neuronais
Tema VIII. Fundamentos de análise sintáctica da linguaxe natural	8.1. Fundamentos da análise sintáctica de constituíntes 8.2. Fundamentos da análise sintáctica de dependencias

Planning				
Methodologies / tests	Competencies	Ordinary class hours	Student?s personal work hours	Total hours
Guest lecture / keynote speech	A28 B4 B8 B9 C4	21	27.3	48.3



Objective test	A28 B2 B3 B7 B9	3	18	21
Laboratory practice	A28 B2 B3 B4 B7 B8 B9 B10 C1	21	35.7	56.7
Personalized attention		24	0	24

(*)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	Nas clases teóricas, os profesores desenvolverán o contido teórico e algúns exercicios complementarios
Objective test	Proba na que se evaluarán os coñecementos adquiridos
Laboratory practice	Permiten ao alumno aprender con eficacia a través de actividades aplicadas, neste caso realizando proxectos en forma de prácticas.

Personalized attention	
Methodologies	Description
Laboratory practice Guest lecture / keynote speech	O desenvolvemento das clases farase de acordo co progreso dos alumnos nas capacidades de comprensión e asimilación dos contidos ensinados, combinando o progreso xeral da clase cunha atención específica aos alumnos que presentan grandes dificultades na tarefa de aprendizaxe e con un apoio adicional a aqueles que presentan maior facilidade e queren ampliar o coñecemento. Dada a natureza personalizada das titorías, estas non deberían dedicarse a ampliar o contido con novos conceptos, senón para aclarar os conceptos xa discutidos. O profesor tamén debe usalos como unha interacción que lle permita sacar conclusións sobre o grao de asimilación da materia polos estudantes.

Assessment			
Methodologies	Competencies	Description	Qualification
Laboratory practice	A28 B2 B3 B4 B7 B8 B9 B10 C1	Realización segundo as condicións establecidas no enunciado de cada práctica.	40
Objective test	A28 B2 B3 B7 B9	Realizarse un exame ao final do semestre e a correspondente 2a oportunidade. Este exame centrarase principalmente nos aspectos teóricos e exercicios asociados. Non obstante, poderá preguntarse algunha cuestión relacionada coas prácticas e os traballos tutelados realizados.	60

Assessment comments



O exame computa o 60% da nota e o 40% restante repártese entre as prácticas e calquera outra actividade avaliable realizada ao longo do curso. Será necesario acadar un mínimo de 3 (sobre 6, no exame) e un 2 (sobre 4, no conxunto das prácticas e demais actividades). En caso de non acadar algún dos mínimos, a máxima nota que se poderá obter é un 4.5.

Se as prácticas ou outras actividades se levan a cabo en grupos, todos os membros do grupo serán responsables solidariamente polo traballo realizado e entregado así como as súas posibles implicacións. Para aprobar a materia é necesario aprobar todas e cada unha das seccións da avaliación. Para a segunda oportunidade, manteranse os resultados obtidos en cada sección (exame e prácticas) da primeira oportunidade.

No caso de estudantes a tempo parcial, a falta de asistencia ás clases de prácticas e TGR xustificadas non será penalizada.

O estudante que non concorra á proba obxectiva no período oficial de avaliación, terá a condición de ?Non presentado? (NP).

A realización fraudulenta das probas ou actividades de avaliación, unha vez comprobada, implicará directamente a cualificación de suspenso '0' na materia na oportunidade correspondente.

Sources of information

Basic	- Aho, A.V.; Lam, M.; Sethi, R.; Ullman, J.D. (2008). Compiladores: Principios, técnicas y herramientas. Addison-Wesley - Hopcroft, J.E.; Motwani, R.; Ullman, J. D. (2002). Introducción a la teoría de autómatas, lenguajes y computación. Addison-Wesley - Garrido, A.; Iñesta, J.M.; Moreno, F.; Pérez, J.A. (2004). Diseño de compiladores. Publicaciones de la Universidad de Alicante - Daniel Jurafsky; James H. Martin (2022). Speech and Language Processing, Draft 3rd Edition. https://web.stanford.edu/~jurafsky/slp3/ - Christopher D. Manning; Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing. MIT Press (ISBN 78-0262133609)
Complementary	- Sudkamp, T.A. (1994). Languages and machines. Addison-Wesley - Aho, A.V.; Ullman, J.D. (1972). The theory of parsing, translation and compiling. Prentice-Hall - Loudon, D.K. (2004). Construcción de compiladores. Principios y Práctica. Paraninfo Thomson Learning - Yoav Goldberg (2017). Neural Network Methods for Natural Language Processing. Morgan & Claypool Publishers (ISBN 978-1627052986) - Lewis Tunstall, Leandro von Werra, and Thomas Wolf (2022). Natural Language Processing with Transformers. O'Reilly Media (ISBN 9781098103248) - Yue Zhang and Zhiyang Teng (2021). Natural Language Processing: A Machine Learning Perspective. Cambridge University Press (ISBN 9781108420211)

Recommendations

Subjects that it is recommended to have taken before

Linear Algebra/614G02001
Fundamentals of Programming II/614G02009
Fundamentals of Programming I/614G02004
Statistical Inference/614G02007
Probability and Basic Statistics/614G02003

Subjects that are recommended to be taken simultaneously

Subjects that continue the syllabus

Natural Language Processing and Text Mining/614G02043

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.