Guía Docente					
Datos Identificativos					2012/13
Asignatura (*)	Extracción e Recuperación da Información			Código	614434003
Titulación	Mestrado Universitario en Computación				
Descriptores					
Ciclo	Período	Curso		Tipo	Créditos
Mestrado Oficial	1º cuadrimestre	Primeiro		Obrigatoria	6
Idioma	CastelánGalegoInglés				
Prerrequisitos					
Departamento	Computación				
Coordinación	Barreiro Garcia, Álvaro		Correo electrónico	alvaro.barreiro@udc.es	
Profesorado	Barreiro Garcia, Álvaro		Correo electrónico	alvaro.barreiro@udc.es	
	Blanco Gonzalez, Roi			roi.blanco@udc.es	
	Vilares Ferro, Jesus			jesus.vilares@udc.es	
Web	campusvirtual.udc.es				

Descrición xeral	DESCRIPCIÓN Esta asignatura aborda la extracción y recuperación de información sobre repositorios de documentos textuales y sobre la web. En el caso de la Recuperación de Información (RI), se estudiarán modelos, técnicas y algoritmos que permiten la recopilación, indexación y búsqueda de documentos sobre colecciones de textos que van desde el orden de megabytes y gigabytes, como el caso de los repositorios de documentos, hasta el orden de terabytes, como en el caso de la web. Por su parte los procesos de Extracción de Información (EI) toman como entrada textos no estructurados y escritos en lenguaje natural, y obtienen a la salida datos no ambiguos representados en un formato definido previamente por el usuario. Los datos extraídos pueden bien ser mostrados directamente al usuario, bien ser almacenados para un análisis posterior, o bien ser utilizados a su vez por los mecanismos de indexación de las aplicaciones de recuperación de información. Las técnicas y algoritmos empleados en tales procesos serán también objeto de estudio en la asignatura. De este modo mientras que un sistema de RI localiza los documentos relevantes al usuario y los presenta al mismo, los sistemas de EI, previa la especificación de los mismos, buscan reconocer las entidades, localizaciones o eventos referidos en dichos textos, permitiendo así presentarle al usuario aquella información específica en la cual está interesado. Sin embargo, si bien los procesos de RI y EI facilitan el tratamiento automático de grandes cantidades de información, ninguno de ellos es capaz de facilitar respuestas precisas a preguntas concretas y arbitrarias formuladas por los usuarios ya que dichos sistemas no han sido concebidos para ello. Mientras los sistemas de RI únicamente permiten devolver una lista de documentos supuestamente relevantes con respecto al contenido de la pregunta, debiendo ser el usuario quien la busque, los sistemas de EI no permiten el tratamiento de preguntas arbitrarias, ya que el tipo de información requerida debe haber sido definida por anticipado. Es por ello que surge un tercer tipo de sistemas, los de Búsqueda de Respuestas (BR), los cuales, haciendo uso combinado de técnicas de RI y EI, permiten localizar respuestas concretas en grandes volúmenes de documentos. El estudio de dichos sistemas conforma la parte final de la asignatura. CONTEXTUALIZACIÓN En el contexto del Master Universitario de Computación, esta materia permite al estudiante ampliar su formación en el ámbito del procesamiento automático de la información, hasta ahora centrado en casos clásicos de procesamiento de los datos almacenados en registros o bases de datos, para dar paso al tratamiento inteligente de la información contenida en los propios textos y al concepto de biblioteca digital. Asimismo, por la propia complejidad de este tipo de sistemas, el curso de esta asignatura permitirá al alumno adquirir conocimientos que le ayudarán a diseñar e implementar otros sistemas de naturaleza compleja similar. Por otra parte, los temas de contenido metodológico le permitirán abordar los aspectos básicos del diseño de experimentos y evaluación, valiosos de por sí tanto para la realización de una tesis doctoral en estos temas como para cualquier ámbito de las ciencias de la computación que incluya componentes experimentales.

Competencias da titulación	
Código	Competencias da titulación
A1	Adquirir coñecementos de Lóxicas Computacionais e as súas principais aplicacións a outras áreas específicas de investigación en Computación tales como Raonamento Automático, Representación do Coñemento, Razoamento Temporal e Espacial, Sistemas Multiaxente, Web semántica, Verificación Formal, etc.
A2	Comprender os conceptos básicos da aprendizaxe computacional, as diferentes técnicas dispoñibles e o seu ámbito de aplicabilidade. Ser capaz de aplicar as distintas técnicas de aprendizaxe empregando unha metodoloxía axeitada.

A3	Coñecemento dos principais aspectos de modelado formal e de avaliación do rendemento dos Sistemas Distribuídos e Concorrentes.
A4	Posuír unha ampla comprensión dos sistemas de Xestión da Información, desde os aspectos máis técnicos como as Estruturas de Datos Compactas e os correspondentes algoritmos de uso, ata as máis avanzadas técnicas de Recuperación da Información, Extracción de Información e Procura de Respostas.
B1	Ser capaz de formular xuízos a partir dunha información que, sendo incompleta ou limitada, inclúa reflexións sobre as responsabilidades sociais e éticas vinculadas á aplicación dos seus coñecementos e xuízos.
B2	Destreza na adquisición do coñecemento, análise do estado da arte e bibliografía relevante nunha área de investigación.
B3	Capacidade para identificar problemas e formular adecuadamente as hipóteses a contrastar seguindo unha metodoloxía científica.
B4	Aplicación do método científico mediante análise empírico das hipóteses formuladas ou mediante demostración formal, no caso de propiedades matemáticas. Destreza no deseño de experimentos e a análise de resultados.
B7	Acostumarse ó uso do inglés como principal idioma de adquisición e transmisión de coñecemento científico e de investigación.
B8	Coñecer resultados recentes en áreas de investigación punteiras e presentados de primeira man polos seus propios autores ou especialistas de recoñecido prestixio.
C1	Expresarse correctamente, tanto de forma oral coma escrita, nas linguas oficiais da comunidade autónoma.
C2	Dominar a expresión e a comprensión de forma oral e escrita dun idioma estranxeiro.
C3	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	Desenvolverse para o exercicio dunha cidadanía aberta, culta, crítica, comprometida, democrática e solidaria, capaz de analizar a realidade, diagnosticar problemas, formular e implantar solucións baseadas no coñecemento e orientadas ao ben común.
C5	Entender a importancia da cultura emprendedora e coñecer os medios ao alcance das persoas emprendedoras.
C6	Valorar criticamente o coñecemento, a tecnoloxía e a información dispoñible para resolver os problemas cos que deben enfrontarse.
C7	Asumir como profesional e cidadán a importancia da aprendizaxe ao longo da vida.
C8	Valorar a importancia que ten a investigación, a innovación e o desenvolvemento tecnolóxico no avance socioeconómico e cultural da sociedade.

Resultados da aprendizaxe			
Competencias de materia (Resultados de aprendizaxe)	Competencias da titulación		
Conocer, comprender y analizar los distintos modelos de Recuperación de Información (RI), Extracción de Información (EI) y Búsqueda de Respuestas (BR), así como las técnicas para su implementación eficiente y la metodología de evaluación de los mismos.	AI1 AI2 AI3 AI4	BI2 BI7 BI8	CM2 CM6
Conocer, comprender y analizar las plataformas software para la creación de estos sistemas.	AI3 AI4	BI2 BI7 BI8	CM2 CM6
Diseñar y construir nuevos sistemas de RI, EI y BR o introducir mejoras en sistemas existentes.	AI1 AI2 AI3 AI4	BI2 BI3 BI7 BI8	CM6 CM7 CM8
Usar las técnicas y métodos propuestos en la asignatura para resolver problemas reales de procesamiento inteligente de la información.	AI1 AI2 AI3 AI4	BI2 BI3 BI4	CM3 CM6 CM8
Asumir la complejidad del lenguaje humano y las limitaciones existentes para su tratamiento automático por ordenador.	AI1 AI2 AI4	BI3	CM1 CM2 CM6
Reconocer los fenómenos lingüísticos que son tratables en los procesos de RI, EI y BR y aquéllos que no lo son, así como conocer y comprender las técnicas para su tratamiento.	AI1 AI2 AI4	BI3	CM1 CM2

Buscar soluciones parciales a un problema ante la imposibilidad de obtener soluciones generales.	AI2 AI4	BI3	CM6
Planear y realizar la evaluación de dichos sistemas. Analizar los resultados obtenidos para mejorarlos en su eficacia y eficiencia.	AI4	BI3 BI4	
Ser capaces de un correcto tratamiento de los aspectos éticos, de privacidad, confidencialidad y de seguridad de tales sistemas.	AI4	BI1	CM6
Valorar el esfuerzo que requiere realizar avances en un campo complejo.			CM4 CM5 CM6 CM7 CM8
Reconocer el esfuerzo y las aportaciones de las comunidades de desarrollo de software y de las comunidades de investigación en la creación y recopilación de recursos en las áreas de RI, EI y BR.			CM4 CM5 CM6 CM7 CM8

Contidos	
Temas	**Subtemas**
INTRODUCCIÓN A LA RECUPERACIÓN DE INFORMACIÓN (RI)	Modelo booleano de recuperación de información. Documentos, términos, vocabulario. Recuperación de información tolerante.
MODELO DE ESPACIO VECTORIAL DE RI	Representación de docuementos, consultas y medidas de similaridad. Esquemas de pesado. Normalización. Implementación eficiente.
MODELO CLASICO PROBABILISTICO DE RI	Probability Ranking Principle Derivación del modelo clásico probabilístico. Otros modelos probabilísticos: 2-Poisson, Okapi, Redes de Inferencia. Implementación eficiente.
MODELO ESTADÍSTICO DE LENGUAJE DE RI	Modelos de lenguaje. Suavización. Aprendizaje y estimación de parámetros. Modelos de lengujaje basados en relevancia. Implementación eficiente.
MODELO LATENT SEMANTIC INDEXING (LSI)	Reducción de dimensionalidad basada en SVD. Derivación del modelo LSI. Cuestiones sobre la escalabilidad del modelo y nuevas aproxiamciones: LSI eficiente, LPI, etc.
EVALUACIÓN EN RI	Tareas y métricas. Colecciones de referencia. TREC, WEB, BLOGS Significancia estadística.
REALIMENTACIÓN DE RELEVANCIA, CLUSTERING Y CLASIFICACIÓN	Realimentación de relevancia bajo el modelo vectorial (Rocchio) y probabilístico. Local Contex Analisys (LCA) y explansión de consultas. Clustering de documentos. Clasificación de documentos.
CONSTRUCCIÓN Y COMPRESIÓN DE INDICES. PROCESAMIENTO DE QUERIES	Algoritmos de construcción de índices. Algoritmos de compresión de índices: compresión de listas, compresión de frecuencias, compresión del léxico. Procesamiento eficiente de consultas.

RI WEB	Modelos de retrieval para el web. Análisis de links. Page Rank y HITS. Implementación de search engines. Oportunidades de RI en el web.
RI PARALELA Y DISTRIBUIDA	RI paralela y distribuida. Modelos de RI distribuida: selección de recursos, enrutado de consultas, fusión de resultados. Aplicaciones novedoes en RI distribuida.
PROCESAMIENTO DEL LENGUAJE NATURAL (PLN) EN RI	Variación lingüística. Tratamiento de la variación morfológica. Stemming. Tratamiento de la variación léxico-semántica. WordNet y EuroWordNet. Tratamiento de la variación sintáctica.
RI MULTILINGÜE E INTERLINGÜE	Impacto del multilingüismo sobre la RI. Aproximaciones al problema del multilingüismo. Traducción Automática (TA): conceptos básicos y problemática. Aproximaciones a la TA: técnicas "clásicas" y técnicas estadísticas. Aplicaciones de la TA en RI Interlingüe. Foros de evaluación: CLEF, NTCIR y FIRE.
EXTRACCIÓN DE INFORMACIÓN (EI)	Conceptos básicos. Arquitectura de un sistema de EI. Tareas de EI. Evaluación en EI. Ejemplos de sistemas de EI: FASTUS y otros.
BÚSQUEDA DE RESPUESTAS (BR)	Conceptos básicos. BR vs. RI/EI. Arquitectura de un sistema de BR. Procesamiento de la pregunta. Recuperación y selección de documentos/pasajes. Extracción de la respuesta. Evaluación en BR.

Planificación			
Metodoloxías / probas	Horas presenciais	Horas non presenciais / traballo autónomo	Horas totais
Sesión maxistral	30	60	90
Traballos tutelados	5	20	25
Lecturas	0	10	10
Proba mixta	3	12	15
Atención personalizada	10	0	10
*Os datos que aparecen na táboa de planificación son de carácter orientativo, considerando a heteroxeneidade do alumnado			

Metodoloxías	
Metodoloxías	Descrición
Sesión maxistral	En las clases presenciales de teoría, el profesor realizará una breve descripción de los contenidos temáticos y de los objetivos básicos perseguidos, con el fin de dotar al alumno de una visión global de la materia. Además tratará de establecer interrelaciones con otros conceptos previamente adquiridos, de forma que se pueda establecer una línea temporal, y expondrá la bibliografía recomendada. Seguidamente pasará a desarrollar los contenidos teóricos, utilizando como método la clase magistral.

Traballos tutelados	Las clases teóricas serán complementadas, además de por lecturas (véase ítem correspondiente), por la elaboración y exposición por parte de los alumnos de trabajos sobre temas de caractér más específico y/o aplicado que los que componen el grueso de los contenidos de las clases magistrales.
Lecturas	Lecturas de fuentes bibliográficas de interés para la ampliación y consolidación de los conocimientos expuestos en las clases magistrales y seminarios.
Proba mixta	Evaluación escrita de los contenidos fundamentales expuestos en las clases magistrales.

Atención personalizada	
Metodoloxías	**Descrición**
Traballos tutelados Lecturas Sesión maxistral	La labor del profesor será la de supervisar el trabajo y formación del alumno, solucionando dudas, corrigiendo errores de interpretación, sugiriendo lecturas, etc., no sólo como grupo, sino también como individuo.

Avaliación		
Metodoloxías	**Descrición**	**Cualificación**
Traballos tutelados	Se valolará la calidad y cobertura de los contenidos del tema asignado, el trabajo desarrollado en su elaboración a partir de las fuentes disponibles, la comprensión y dominio de dichos conocimientos por parte de sus autores, así como la estructura y claridad de la exposición.	40
Proba mixta	Se evaluarán el dominio de conocimientos teóricos adquiridos durante el curso y su aplicación en resolución de problemas.	55
Sesión maxistral	Se valolará la asistencia y participación activa en las clases y tutorías colectivas.	5

Observacións avaliación
El alumno, bien indivualmente, bien por parejas, deberá realizar un trabajo sobre un tema designado por el profesor y dentro del ámbito de la asignatura. Dicho trabajo será expuesto delante del profesor y el resto de los estudiantes para su evaluación y discusión. Habrá también un examen final al término del curso en el que se valorarán tanto los contenidos teóricos adquiridos como la habilidad del alumno para su aplicación en la resolución de problemas prácticos del ámbito. La puntuación asignada a cada una de los apartados del examen irá consignada en la prueba. Asimismo, a la hora de ser evaluado, se valolará positivamente la asistencia y participación activa del estudiante en el marco de la asignatura. El estudio de la asignatura no puede plantearse como una mera actividad de estudio memorístico de las técnicas y algoritmos presentados en clase y de lectura de la bibliografía, sino que deberá plantearse con vistas a su comprensión y análisis de la aplicabilidad práctica de los contenidos presentados en el marco de la asignatura.

Fontes de información

Bibliografía básica	- W. John Hutchings y Harold L. Somers (1992). An Introduction to Machine Translation. Academic Press, Londres/San Diego - Gregory Grefenstette (ed.) (1998). Cross-language information retrieval. Kluwer Academic Publishers, Boston - Christopher D. Manning y Hinrich Schütze (1999). Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge (Massachusetts, EE.UU.)/Londres (Reino Unido) - David A. Grossman y Ophir Frieder (1998). Information Retrieval: Algorithms and Heuristics. Kluwer Academic Publishers - C.D. Manning, P. Raghavan y H. Schutze (2008). Introduction to Information Retrieval. Cambridge. Cambridge University Press - Peter Jackson e Isabelle Moulinier (2007). Natural language processing for online applications : text retrieval, extraction and categorization (2nd ed.). John Benjamins, Amsterdam/Philadelphia - Marius Pasca (2003). Open-domain question answering from large text collections. CSLI Publications, Standford - W.B. Croft, D. Metzler y T. Strohman (2009). Search Engines. Information Retrieval in Practice. Pearson Education - Daniel Jurafsky y James H. Martin (2009). Speech and Language Processing. An Introduction to Natural Language Processing, Computational Linguistics and Speech Recognition (2nd ed.). Pearson Prentice Hall, Upper Saddle River, New Jersey, EE.UU - Jerry R. Hobbs (1993). The generic information extraction system. En Proceedings of the 5th Conference on Message understanding (MUC-5), pág. 87-91. Morgan Kauffman Publishers, San Francisco, USA - E. Voorhees and D.K. Harman (2005). TREC: experiment and evaluation in information retrieval. MIT Press
Bibliografía complementaria	- A. Moffat y A. Turpin (2002). Compression and Coding Algorithms. Kluwer Academic Publishers - Fotis Lazarinis, Jesús Vilares, John I. Tait, J. & Efthimis N. Efthimiadis (2009). Current research issues and trends in non-English Web searching. En Special Issue on Non-English Web Retrieval, Journal of Information Retrieval, 12(3), 230-250. Springer , Berlin - Heidelberg- New York - Piek Vossen (ed.) (1998). EuroWordNet. A Multilingual Database with Lexical Semantic Networks. Kluwer Academic Publishers - J.R. Hobbs, D. Appelt, J. Bear, D. Israel, M. Kameyama, M. Stickel y M. Tyson (1997). FASTUS - A Cascaded Finite-State Transducer for Extracting Information from Natural-Language Text. Ch. 13 of Finite-State Language Processing. MIT Press - R. K. Belew (2001). Finding Out About. Cambridge Press - Robert Dale, Hermann Moisi y Harold Somers (eds.) (2000). Handbook of Natural Language Processing. Marcel Dekker, Inc., Nueva York/Basilea - M. Constantino y P. Coletti (2008). Information Extraction in Finance. WIT Press, Southampton, UK - Marie-Francine Moens (2006). Information Extraction: Algorithms and Prospects in a Retrieval Context. Springer , Berlin - Heidelberg- New York - C. J. Van Rijsbergen (1979). Information Retrieval (2nd ed.). Butterworths, London - W.B. Croft y J. Lafferty (2003). Language Modeling for Information Retrieval. Kluwer Academic Publishers - A. Arampatzis, Th. P. van der Weide, P. van Bommel y C.H.A. Koster (2000). Linguistically-motivated Information Retrieval. En Vol. 69 de Encyclopedia of Library and Information Science, pág. 201-222. Marcel Dekker - H. Witten, A. Moffat, y T. C. Bell (1999). Managing Gigabytes: Compressing and Indexing Documents and Images (2nd ed.). Morgan Kaufmann - R. Baeza-Yates y B. Ribeiro-Neto (1999). Modern Information Retrieval. Addison Wesley - K. Kishida (2005). Technical issues of cross-language information retrieval: a review. En Special Issue on Cross-Language Information Retrieval, Information Processing & Management, 41(3), 433-455. Elsevier

Recomendacións
Materias que se recomenda ter cursado previamente

Materias que se recomenda cursar simultaneamente
Procesamento Avanzado da Linguaxe Natural/614434011
Materias que continúan o temario
Observacións

(*)A Guía docente é o documento onde se visualiza a proposta académica da UDC. Este documento é público e non se pode modificar, salvo casos excepcionais baixo a revisión do órgano competente dacordo coa normativa vixente que establece o proceso de elaboración de guías