



Guía Docente

Datos Identificativos					2012/13
Asignatura (*)	Extracción e Recuperación da Información		Código	614434003	
Titulación					
Descritores					
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 Blanco Gonzalez, Roi Vilares Ferro, Jesus		Correo electrónico	alvaro.barreiro@udc.es roi.blanco@udc.es jesus.vilares@udc.es	
Web	campusvirtual.udc.es				



Descrición xeral	<p>DESCRIPCIÓN</p> <p>Esta asignatura aborda a extracción e recuperación de información sobre repositorios de documentos textuais e sobre a web.</p> <p>En o caso da Recuperación de Información (RI), se estudarán modelos, técnicas e algoritmos que permiten a recopilación, indexación e búsqueda de documentos sobre coleccións de textos que van desde o orden de megabytes e gigabytes, como o caso dos repositorios de documentos, ata o orden de terabytes, como en o caso da web.</p> <p>Por a súa parte os procesos de Extracción de Información (EI) toman como entrada textos non estruturados e escritos en lingua natural, e obtíñen a saída datos non ambiguos representados en un formato definido previamente polo usuario. Os datos extraídos poden ben ser mostrados directamente ao usuario, ben ser almacenados para un análise posterior, ou ben ser utilizados a súa vez polos mecanismos de indexación das aplicacións de recuperación de información. As técnicas e algoritmos empregados en tales procesos serán tamén obxecto de estudo na asignatura.</p> <p>De este modo mentres que un sistema de RI localiza os documentos relevantes ao usuario e llos presenta ao mesmo, os sistemas de EI, previa a especificación dos mesmos, buscan reconecer as entidades, localizacións ou eventos referidos en dichos textos, permitindo así presentarlle ao usuario aquela información específica na cal está interesado.</p> <p>Sin embargo, si ben os procesos de RI e EI facilitan o tratamento automático de grandes cantidades de información, ningunha deles é capaz de facilitar respostas precisas a preguntas concretas e arbitrarias formuladas polos usuarios xa que dichos sistemas non han sido concebidos para el. Mentres os sistemas de RI unicamente permiten devolver unha lista de documentos supuestamente relevantes con respecto ao contido da pregunta, debendo ser o usuario quen a busque, os sistemas de EI non permiten o tratamento de preguntas arbitrarias, xa que o tipo de información requirida debe haber sido definida por anticipado. Es por el que surge un terceiro tipo de sistemas, os de Búsqueda de Respostas (BR), os cales, facendo uso combinado de técnicas de RI e EI, permiten localizar respostas concretas en grandes volúmenes de documentos. O estudo de dichos sistemas conforma a parte final da asignatura.</p> <p>CONTEXTUALIZACIÓN</p> <p>En o contexto do Master Universitario de Computación, esta materia permite ao estudante ampliar a súa formación en o ámbito do procesamento automático da información, ata agora centrado en casos clásicos de procesamento dos datos almacenados en registros ou bases de datos, para dar paso ao tratamento intelixente da información contida en os propios textos e ao concepto de biblioteca dixital.</p> <p>Asimismo, por a propia complexidade deste tipo de sistemas, o curso de esta asignatura permitirá ao alumno adquirir coñecementos que le ayudarán a deseñar e implementar outros sistemas de natureza compleja similar.</p> <p>Por outra parte, os temas de contido metodolóxico le permitirán abordar os aspectos básicos do deseño de experimentos e avaliación, valiosos de por si tanto para a realización de unha tese doutoral en estes temas como para calquera ámbito das ciencias da computación que incluya componentes experimentais.</p>
-------------------------	---

Competencias da titulación	
Código	Competencias da titulación

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 documentos, 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 lenguaje 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 aproximaciones: 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 Context Analysis (LCA) y expansión 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 novedosas 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)	<p>Conceptos básicos.</p> <p>BR vs. RI/EI.</p> <p>Arquitectura de un sistema de BR.</p> <p>Procesamiento de la pregunta.</p> <p>Recuperación y selección de documentos/pasajes.</p> <p>Extracción de la respuesta.</p> <p>Evaluación en BR.</p>
-----------------------------	---

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 presenciais 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 carácter 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 valorará 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 valorará la asistencia y participación activa en las clases y tutorías colectivas.	5

Observacións avaliación



El alumno, bien individualmente, 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.

Habrà 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 valorará 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. Schütze (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	<ul style="list-style-type: none"> - A. Moffat y A. Turpin (2002). Compression and Coding Algorithms. Kluwer Academic Publishers - Fotis Lazarinis, Jesús Vilares, John I. Tait, J. & Eftimis N. Eftimiadis (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