Teaching Guide				
Identifying Data				2012/13
Subject (*)	Recuperación da información e web semántica		Code	614502010
Study programme	Mestrado Universitario en Enxeñaría Informática (plan 2012)			
Descriptors				
Cycle	Period	Year	Type	Credits
Official Master's Degree	1st four-month period	First	Obligatoria	6
Language	Spanish			
Prerequisites				
Department	ComputaciónTecnoloxías da Información e as Comunicacións			
Coordinador	Barreiro Garcia, Álvaro	E-mail	alvaro.barreiro@udc.es	
Lecturers	Barreiro Garcia, Álvaro Blanco Gonzalez, Roi Cacheda Seijo, Fidel Vázquez Naya, José Manuel	E-mail	alvaro.barreiro@udc.es roi.blanco@udc.es fidel.cacheda@udc.es jose.manuel.vazquez.naya@udc.es	
Web				
General description	Los modelos, técnicas y algoritmos de recuperación de información estudiados en esta materia permitirán a los estudiantes comprender la arquitectura de los Search Engines para el web. Además los contenidos prácticos de la misma les capacitarán para construir sus propios buscadores para trabajar sobre repositorios de documento o la web. Además durante los últimos años ha habido un interés creciente en idear una web semántica a partir de meta-datos y anotaciones. Una web basada en documentos xml y tags, meta-datos y esquemas, sin duda facilitaría los enormes retos a los que se enfrenta la recuperación de información web. En esta asignatura se abordan también los modelos, técnicas y algoritmos de mayor impacto desarrollados en los últimos años con el objetivo de materializar una web semántica. La Recuperación de Información en grandes colecciones de documentos y en la web plantea enormes retos (volumen de datos, datos distribuidos, alto porcentaje de datos volátiles, datos no estructurados y redundantes, heterogeneidad, calidad de los datos y confianza) y la Web Semántica parte ya del gran reto de la extracción de información cuando los meta-datos no son expuestos públicamente y plantea nuevos retos como los del matching de ontologías, resolución de entidades o una dificultad mayor en cuanto a la heterogeneidad y calidad de los datos y a la indexación y búsqueda semántica. Por todo ello la Recuperación de Información y la Web semántica constituyen uno de los campos de mejores salidas profesionales en informática con oportunidades de negocio y empleo no sólo en las grandes compañías de Search Engines sino también en muchas pequeñas y medianas compañías.			

Study programme competences	
Code	Study programme competences
A5	Capacidade de comprender e saber aplicar o funcionamento e organización da internet, as tecnoloxías e protocolos de redes de nova xeración, os modelos de compoñentes, sóftware intermediario e servizos.
A9	Capacidade para deseñar e avaliar sistemas operativos e servidores, e aplicacións e sistemas baseados en computación distribuída.
B1	Capacidade de resolución de problemas.
B3	Capacidade de análise e síntese.
B5	Habilidades de xestión da información.
B7	Preocupación pola calidade.
B9	Capacidade para xerar novas ideas (creatividade).
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.
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

Learning outcomes			
Subject competencies (Learning outcomes)	Study programme competences		
Know, understand and analyze different models of information retrieval and semantic web, techniques for their efficient implementation and their evaluation methodology.	AJ5	BJ3	CJ2 CJ6 CJ8
Know, understand and analyze the software platforms used to create these systems.	AJ5	BJ3	CJ2 CJ3 CJ6 CJ7 CJ8
Design and build new systems or improve the existing ones.	AJ5 AJ9	BJ1 BJ3 BJ5 BJ9	CJ3 CJ5 CJ6 CJ7
Plan and perform the evaluation of information retrieval and semantic web systems. Analyze the evaluation results of the systems in order to improve their efficiency and effectiveness.	AJ5 AJ9	BJ1 BJ5 BJ7	CJ3 CJ5 CJ6 CJ7
Be able to treat correctly the ethical, privacy, confidentiality and security aspects of these systems.			CJ6

Contents	
Topic	Sub-topic
Introduction	Information Retrieval and Search Engine Architecture
Information gathering	Crawling and feeds.
Text and Web page processing	Text pre-processing and parsing. Anchor text and Web link analysis, internationalization.
Indexes and ranking.	Building and compressing indexes. Efficient query processing.
Query formulation and results presentation.	Formulation and query re-writing. Snippets. Results visualization.
Information Retrieval Models.	Boolean, Vector-space, probabilistic, language models.
Evaluation	Evaluation of Information Retrieval Systems. Evaluation campaigns. Efficiency and effectiveness metrics. Evaluation design: training, test, statistical significance. Crowd-sourced evaluation.
Text mining.	Document clustering and classification
Distributed and Social search.	Federated search and distributed search. Blogs, micro-blogs and social networks.
Recommender systems	Collaborative filtering. Models and algorithms for recommendation. Recommender systems.
Introduction to the Semantic Web	Semantic Web. Ontologies, definition, types and examples.
Description and resource querying.	XML, RDF and RDF Schema languages. SPARQL query language. OWL language. Tools for ontology development. Libraries for ontology management. RDF repositories.
Reasoning and rules.	Formal logic and reasoning foundations. Semantic rule representation. Reasoning engines.
Semantic Web applications.	Linked Data, FOAF, Dublin Core, WordNet. Semantic annotations. Semantic Search. Semantic Web services.

Planning			
Methodologies / tests	Ordinary class hours	Student?s personal work hours	Total hours

Lecturas	1	15	16
Prácticas de laboratorio	20	30	50
Solución de problemas	4	12	16
Proba mixta	2	18	20
Sesión maxistral	16	32	48
Personalized attention	0		0
(*)The information in the planning table is for guidance only and does not take into account the heterogeneity of the students.			

Methodologies	
Methodologies	**Description**
Lecturas	Readings in order to consolidate and complement the knowledge and skills acquired.
Prácticas de laboratorio	Labs assignments dealing with development platforms in commercial use (Lucene, Terrier, Nutch, Jena, Protege, Pellet)
Solución de problemas	Problems and short questions to consolidate the contents presented in the master classes.
Proba mixta	Test about the fundamental contents of the subject.
Sesión maxistral	The student will attend to the lectures given by the teacher about the different techniques, models and algorithms related to Information Retrieval and the Wemantic Web. The teacher will employ different levels of abstraction-detail and will guide the student in the fundamental and complementary readings.

Personalized attention	
Methodologies	**Description**
Prácticas de laboratorio Solución de problemas	Control of the development of the labs assignment in the allocated lab hours, and the teacher will pay special attention to the student in particularly difficult problems, if necessary.

Assessment		
Methodologies	**Description**	**Qualification**
Prácticas de laboratorio	Control of the labs assignments and evaluation of the results achieved.	50
Proba mixta	Questions related to the knowledge acquired. Questions that involve reasoning over the knowledge acquired, that involve practical problem-solving on real life issues related to Information Retrieval and the Semantic Web.	50

Assessment comments

Sources of information	
Basic	- Bob DuCharme (2011). Learning SPARQL. O'Reilly - C.D. Manning, P. Raghavan, H. Schutze. (2008). Introduction to Information Retrieval. Cambridge University Press - R. Baeza-Yates and B. Ribeiro-Neto. (2011). Modern Information Retrieval (second edition) . Addison Wesley/Pearson Education - F. Cacheda, J.M. Fernández, J. Huete (eds.) (2011). Recuperación de Información. Un enfoque práctico y multidisciplinar. Ra-Ma - W.B. Croft, D. Metzler, T. Strohman. (2009). Search Engines. Information Retrieval in Practice. Pearson Education - John Hebeler, Matthew Fisher, Ryan Blace, Andrew Perez-Lopez, Mike Dean. (2009). Semantic Web Programming. Wiley

Complementary	

Recommendations
Subjects that it is recommended to have taken before
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.