Guía docente de la asignatura

Asignatura	WEB SEMÁNTICA Y EXTRACCIÓN DE INFORMACIÓN		
Materia	SISTEMAS INTELIGENTES Y BASADOS EN CONOCIMIENTO		
Módulo			
Titulación	MÁSTER EN INGENIERÍA INFORMÁTICA		
Plan	510	**Código**	53181
Periodo de impartición	2º CUATRIMESTRE	**Tipo/Carácter**	OPTATIVA
Nivel/Ciclo	MÁSTER	**Curso**	1º
Créditos ECTS	3 ECTS		
Lengua en que se imparte	CASTELLANO		
Profesor/es responsable/s	MERCEDES MARTÍNEZ GONZÁLEZ, TEODORO CALONGE CANO		
Datos de contacto (E-mail, teléfono…)	TELÉFONO: 983 423000 ext. 5607 / ext. 5603 E-MAIL: mercedes@infor.uva.es, teodoro@infor.uva.es		
Horario de tutorías	Véase www.uva.es → Centros → Campus de Valladolid → Escuela Técnica Superior de Ingeniería Informática → Tutorías		
Departamento	INFORMÁTICA		

1. Situación / Sentido de la Asignatura

1.1 Contextualización

La asignatura se enmarca dentro de la materia “Sistemas Inteligentes y Basados en el Conocimiento” y es de carácter optativo. La asignatura aborda los problemas relacionados con la compartición de semántica en la web (web semántica) y su extracción a partir de la información disponible en fuentes diversas para poder ser representada convenientemente y compartida en la web. Problemas todos ellos relacionados con la explosión de información presente en la web cuyo valor implícito aún no se es capaz de explotar convenientemente. Sin embargo, los avances de los últimos años han trasladado este campo de un ámbito exclusivamente investigador, a la presencia cada vez mayor de empresas dedicadas a producir soluciones tecnológicas en este campo, así como de inversiones realizadas por las grandes empresas dedicadas a la manipulación de información en la web en este tipo de tecnologías y aplicaciones (Freebase, de Google; ...).

1.2 Relación con otras materias

1.3 Prerrequisitos

Aunque no es imprescindible, es conveniente tener un buen conocimiento de modelos de datos y lenguajes de consulta, programación, análisis léxico y sintáctico de lenguajes formales, y, finalmente, fundamentos de las técnicas básicas de reconocimiento de patrones.

2. Competencias

2.1 Generales

Código	Descripción
CG1	*Capacidad para proyectar, calcular y diseñar productos, procesos e instalaciones en todos los ámbitos de la ingeniería informática.*
CET5	*Capacidad para analizar las necesidades de información que se plantean en un entorno y llevar a cabo en todas sus etapas el proceso de construcción de un sistema de información*
CET9	*Capacidad para aplicar métodos matemáticos, estadísticos y de inteligencia artificial para modelar, diseñar y desarrollar aplicaciones, servicios, sistemas inteligentes y sistemas basados en el conocimiento*

2.2 Específicas

Código	Descripción
CA1	*Capacidad para comprender y responder a las necesidades de representación de información, conocimiento y semántica en entornos de compartición masiva de información, tales como la web*
CA2	*Capacidad para analizar las necesidades de extracción de información, y aplicar las técnicas adecuadas para extraer información de fuentes de datos no estructuradas*

3. Objetivos

Código	Descripción
R1	Entender los retos de interoperabilidad que supone la compartición de información en la web
R2	Ser capaz de conocer y utilizar los modelos de datos asociados a la web para compartir datos y semántica
R3	Conocer y saber aplicar las tecnologías (estándares, lenguajes de consulta, otros) asociados a estos modelos de datos
R4	Reconocer las necesidades de extracción de información en un problema real y proponer soluciones tecnológicas capaces

	de resolverla
R5	Seleccionar la técnica de extracción de información más adecuada para cada problema
R6	Combinar diferentes técnicas básicas para obtener sistemas más eficientes de extracción de información

4. Tabla de dedicación del estudiante a la asignatura

ACTIVIDADES PRESENCIALES	HORAS	ACTIVIDADES NO PRESENCIALES	HORAS
Clases teórico-prácticas (T/M)	14	Estudio y trabajo autónomo individual	45
Clases prácticas de aula (A)		Estudio y trabajo autónomo grupal	
Laboratorios (L)	10		
Prácticas externas, clínicas o de campo			
Seminarios (S)	**4**		
Tutorías grupales (TG)			
Evaluación (fuera del periodo oficial de exámenes)	2		
Total presencial	**30**	**Total no presencial**	**45**

5. Bloques temáticos

Bloque 1: Web Semántica

Carga de trabajo en créditos ECTS: 1.5

a. Contextualización y justificación

En este bloque se presentan los problemas abordados en el contexto conocido como "Web Semántica" o Web 3.0, y las soluciones tecnológicas que constituyen la base de los desarrollos actuales en este campo, así como los que se espera tengan amplio impacto en un futuro próximo. El alumno entrará en contacto con los desarrollos y aplicaciones que ya utilizan estas soluciones para ofrecer a sus usuarios mejor acceso a información en la web. Existen así buscadores semánticos que permiten a sus usuarios ir más allá de la búsqueda por cadenas, mashup semánticos que integran información en base a sus propiedades, facilitando así la localización de información dispersa acerca de una fuente, pero liberando también al usuario de la tarea de integrarla, redes que facilitan la relación entre nodos de información en base a propiedades con semántica bien definida, etc. Todas estas soluciones están integrándose en la web actual, y son un pilar determinante de sus posibilidades futuras de ofrecer servicios avanzados a unos usuarios cada vez más exigentes.

b. Objetivos de aprendizaje

Código	Descripción
R1	Entender los retos de interoperabilidad que supone la compartición de información en la web
R2	Ser capaz de conocer y utilizar los modelos de datos asociados a la web para compartir datos y semántica
R3	Conocer y saber aplicar las tecnologías (estándares, lenguajes de consulta, otros) asociados a estos modelos de datos

c. Contenidos

TEMA 1: Introducción al problema de la información en la web y sus soluciones

1.1 Deficiencias actuales del acceso a la información en la web.

1.2 Web de datos: semántica en la web.

1.2 Estudio de aplicaciones reales de web semántica y su valor añadido.

1.3 Perspectivas de desarrollos futuros.

TEMA 2: Tecnologías para la representación, manipulación y compartición de semántica en la Web

2.1 Modelos de datos usados para la representación de semántica en la web. Sus posibilidades y uso en casos reales.

2.2 Acceso a la información semántica: lenguajes de consulta, estrategias para la integración.

2.3 Casos de estudio: datos enlazados, mashup semánticos.

d. Métodos docentes

- Sesiones de aula:
  - Clases magistrales participativas y expositivas
  - Aprendizaje basado en problemas
  - Estudios de caso.
- Laboratorio y Prácticas supervisadas:
  - Resolución de problemas y casos prácticos.
  - Aprendizaje basado en problemas.
  - Aprendizaje cooperativo.
  - Estudios del caso.
  - Método de proyectos.
- Seminarios:
  - Aprendizaje basado en problemas
  - Estudios del caso
  - Aprendizaje cooperativo

e.Plan de trabajo

Semana	Contenido	Actividad presencial	Actividad no presencial
Semana 1	Tema 1	Sesiones en aula y laboratorio: 4h.	
Semana 2	Temas 1 y 2	Sesiones en aula y laboratorio: 4 h.	Estudio y solución de problemas. Estudio de casos.
Semana 3	Tema 2	Sesiones en aula y laboratorio: 4 h.	Estudio y solución de problemas. Estudio de casos.
Semana 4	Tema 2	Seminario (2h)	Seguimiento del trabajo práctico del bloque 1.
Semana 5			
Semana 6			
Semana 7			

Semana 8		Seminario (1h)	Seguimiento del trabajo práctico del bloque 1.

f. Evaluación

INSTRUMENTO/PROCEDIMIENTO	PESO EN LA NOTA FINAL	OBSERVACIONES
Entrega práctica	40%	La práctica consistirá en un proyecto, que se irá revisando en sucesivos hitos hasta llegar a la entrega final.
Valoración del trabajo realizado en el laboratorio	5%	Cada tema tiene una serie de ejercicios que los alumnos deben realizar para comprobar su nivel de comprensión y evolución.
Participación en las actividades	5%	

g. Bibliografía básica

- Grigoris Antoniou, Paul Groth, Fran Van Harmelen, Rinke Hoesktra. *A Semantic Web Primer*. 3ª ed. MIT Press. 2012.
- Liyang Yu. *A Developer's Guide to the Semantic Web*. Springer Verlag. 2011.
- Tom Heath and Christian Bizer. *Linked Data: Evolving the Web into a Global Data Space* (1st edition). Synthesis Lectures on the Semantic Web: Theory and Technology, 1:1, 1-136. Morgan & Claypool. 2011.

.

h. Bibliografía complementaria

Sitios web dedicados a web semántica, linked data, etc.

i. Recursos necesarios

Libros de texto, presentaciones audiovisuales, material disponible en el aula virtual de la asignatura.

Bloque 2: Extracción de Información

Carga de trabajo en créditos ECTS: 1.5

a. Contextualización y justificación

Debido a la ingente cantidad de datos generados y mostrados en la web, se hace preciso establecer una extracción de semántica, para su posterior asimilación y publicación. Para ello, se recurre a técnicas nuevas, pero, muchas de ellas, procedentes de aproximaciones ya conocidas, como la Teoría de Autómatas y Lenguajes Formales, con el objetivo de obtener una estructura semántica, aunque incompleta, pero que sirva de base para posteriores refinamientos. Existen ejemplos de trabajos en esta línea, como los que se pueden derivar de la DBpedia. A partir de esta fase, se aplicarían técnicas de reconocimiento de patrones, que si bien podrían haberse planteado desde un principio, la amplia experiencia en este campo aconseja el uso de otras técnicas que sirvan de "preprocesado" (extracción semántica), en aras de un mejor rendimiento del sistema global.

b. Objetivos de aprendizaje

Código	Descripción
R4	Reconocer las necesidades de extracción de información en un problema real y proponer soluciones tecnológicas capaces de resolverla
R5	Seleccionar la técnica de extracción de información más adecuada para cada problema
R6	Combinar diferentes técnicas básicas para obtener sistemas más eficientes de extracción de información

c. Contenidos

TEMA 3: Introducción al tratamiento de cadenas:

3.1 Definiciones: alfabeto, cadenas, lenguajes, términos, distancia entre cadenas.

3.2 Representación de lenguajes: autómatas y gramáticas.

3.3 Representación de distribuciones de cadenas mediante autómatas y gramáticas.

- 3.3.1 Autómatas probabilísticos
- 3.3.2 Gramáticas independientes del contexto probabilísticas
- 3.3.3 Distancia entre gramáticas
- 3.3.4 Distancia entre distribuciones de cadenas

TEMA 4: Técnicas de aprendizaje automático de lenguajes y su aplicación en la extracción de informacón

4.1 Identificación de lenguajes.

4.2 Aprendizaje a partir de texto.

4.3. Aprendizaje activo aplicado a la Web Semántica

4.4 Aprendizaje de distribuciones de cadenas.

4.5 Aprendizaje de texto corregido mediante aproximaciones probabilísticas.

d. Métodos docentes

- Sesiones de aula:
  - Clases magistrales participativas y expositivas
  - Aprendizaje basado en problemas
  - Estudios de caso.
- Laboratorio y Prácticas supervisadas:
  - Resolución de problemas y casos prácticos.
  - Aprendizaje basado en problemas.
  - Aprendizaje cooperativo.
  - Estudios del caso.
  - Método de proyectos.
- Seminarios:
  - Aprendizaje basado en problemas
  - Estudios del caso
  - Aprendizaje cooperativo

e. Plan de trabajo

Semana	Contenido	Actividad presencial	Actividad no presencial
Semana 1			
Semana 2			
Semana 3			
Semana 4	Tema 3	Sesiones en aula y laboratorio: 2h.	Estudio y solución de problemas. Estudio de casos.
Semana 5	Tema 3	Sesiones en aula y laboratorio: 4h	Estudio y solución de problemas. Estudio de casos.
Semana 6	Tema 4	Sesiones en aula y laboratorio: 4h	Estudio y solución de problemas. Estudio de casos.
Semana 7	Tema 4	Sesiones en aula y laboratorio: 4h	Estudio y solución de problemas. Estudio de casos.
Semana 8	Tema 4	Seminario (1h) Evaluación (2h)	Seguimiento del trabajo práctico del bloque 2.

f. Evaluación

INSTRUMENTO/PROCEDIMIENTO	PESO EN LA NOTA FINAL	OBSERVACIONES
Entrega práctica	40%	La práctica consistirá en un proyecto, que se irá revisando en sucesivos hitos hasta llegar a la entrega final.
Valoración del trabajo realizado en el laboratorio	5%	Cada tema tiene una serie de ejercicios que los alumnos deben realizar para comprobar su nivel de comprensión y evolución.
Participación en las actividades	5%	

g. Bibliografía básica

- Colin de la Higuera. *Grammatical Inference: Learning Automata and Grammars*. Ed. Cambridge University Press, 2010.
- Poibeau, T.; Saggion, H.; Piskorski, J.; Yangarber, R. *Multi-source, Multilingual Information Extraction and Summarization*. Series: Theory and Applications of Natural Language Processing. Ed. Springer-Verlag, 2013

h. Bibliografía complementaria

Sitios web dedicados a web semántica, linked data, etc.

i. Recursos necesarios

Libros de texto, presentaciones audiovisuales, material disponible en el aula virtual de la asignatura. Laboratorio con ordenadores personales dotados de analizadores léxicos y sintácticos, como lex y yacc, por ejemplo. Asimismo, se precisa acceso a herramientas, donde estén implementadas técnicas de aprendizaje automático, como Weka, MatlLab u Octave.

6. Temporalización (por bloques temáticos)

BLOQUE TEMÁTICO	CARGA ECTS	PERIODO PREVISTO DE DESARROLLO
Bloque 1: Web Semántica	1.5 ECTS	Semanas 1 a 4
Bloque 2: Extracción de Información	1.5 ECTS	Semanas 4 a 8

7. Sistema de calificaciones – Tabla resumen

INSTRUMENTO/PROCEDIMIENTO	PESO EN LA NOTA FINAL	OBSERVACIONES
Entrega práctica bloque I	40%	La práctica consistirá en un proyecto, que se irá revisando en sucesivos hitos hasta llegar a la entrega final.
Entrega práctica bloque II	40%	La práctica consistirá en un proyecto, que se irá revisando en sucesivos hitos hasta llegar a la entrega final.
Valoración del trabajo realizado en el laboratorio	10%	Cada tema tiene una serie de ejercicios que los alumnos deben realizar para comprobar su nivel de comprensión y evolución.
Participación en las actividades	10%	

CRITERIOS DE CALIFICACIÓN
• **Convocatoria ordinaria:** Los criterios de calificación son los que resultan de aplicar el sistema de calificaciones indicado en el apartado 7 • **Convocatoria extraordinaria:** Mismo sistema que en la convocatoria ordinaria.