PLAN DOCENTE DE LA ASIGNATURA

Curso académico: 2024/2025

Identificación y características de la asignatura				
Código	501326		Créditos ECTS	6
Denominación (español)	Recuperación de la Información y Búsqueda en la Web			
Denominación (inglés)	*Information Retrieval and Web Search*			
Titulaciones	Grado en Ingeniería Informática en Ingeniería del Software			
Centro	Escuela Politécnica			
Semestre	8	Carácter	Optativa	
Módulo	Optatividad en Ingeniería del Software			
Materia	Ingeniería Multimedia			
Profesor/es				
Nombre	Despacho		Correo-e	Página web
Félix Rodríguez Rodríguez	23 (Edif.Telecos)		felixr@unex.es	Web EPCC (felixr) y madiba.unex.es
Área de conocimiento	Lenguajes y Sistemas Informáticos			
Departamento	Ingeniería de Sistemas Informáticos y Telemáticos			
Profesor coordinador (si hay más de uno)				

Competencias

Competencias básicas:

- **CB1**: Que los estudiantes hayan demostrado poseer y comprender conocimientos en un área de estudio que parte de la base de la educación secundaria general, y se suele encontrar a un nivel que, si bien se apoya en libros de texto avanzados, incluye también algunos aspectos que implican conocimientos procedentes de la vanguardia de su campo de estudio.
- **CB2**: Que los estudiantes sepan aplicar sus conocimientos a su trabajo o vocación de una forma profesional y posean las competencias que suelen demostrarse por medio de la elaboración y defensa de argumentos y la resolución de problemas dentro de su área de estudio.
- **CB3**: Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética.
- **CB4**: Que los estudiantes puedan transmitir información, ideas, problemas y soluciones a un público tanto especializado como no especializado.
- **CB5**: Que los estudiantes hayan desarrollado aquellas habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía.

Competencias generales:

- **CG01**: Capacidad para concebir, redactar, organizar, planificar, desarrollar y firmar proyectos en el ámbito de la ingeniería en informática que tengan por objeto, de acuerdo con los conocimientos adquiridos para la tecnología específica de Ingeniería

del Software, la concepción, el desarrollo o la explotación de sistemas, servicios y aplicaciones informáticas.

- **CG02**: Capacidad para dirigir las actividades objeto de los proyectos del ámbito de la Informática.
- **CG03**: Capacidad para diseñar, desarrollar, evaluar y asegurar la accesibilidad, ergonomía, usabilidad y seguridad de los sistemas, servicios y aplicaciones informáticas, así como de la información que gestionan.
- **CG04**: Capacidad para definir, evaluar y seleccionar plataformas hardware y software para el desarrollo y la ejecución de sistemas, servicios y aplicaciones informáticas, de acuerdo con los conocimientos adquiridos para la tecnología específica de Ingeniería del Software.
- **CG05**: Capacidad para concebir, desarrollar y mantener sistemas, servicios y aplicaciones informáticas empleando los métodos de la ingeniería del software como instrumento para el aseguramiento de su calidad, de acuerdo con los conocimientos adquiridos para la tecnología específica de Ingeniería del Software.
- **CG06**: Capacidad para concebir y desarrollar sistemas o arquitecturas informáticas centralizadas o distribuidas integrando hardware, software y redes.
- **CG07**: Capacidad para conocer, comprender y aplicar la legislación necesaria durante el desarrollo de la profesión de Ingeniero Técnico en Informática y manejar especificaciones, reglamentos y normas de obligado cumplimiento.
- **CG08**: Conocimiento de las materias básicas y tecnologías, que capaciten para el aprendizaje y desarrollo de nuevos métodos y tecnologías, así como las que les doten de una gran versatilidad para adaptarse a nuevas situaciones.
- **CG09**: Capacidad para resolver problemas con iniciativa, toma de decisiones, autonomía y creatividad. Capacidad para saber comunicar y transmitir los conocimientos, habilidades y destrezas de la profesión de Ingeniero Técnico en Informática.
- **CG10**: Conocimientos para la realización de mediciones, cálculos, valoraciones, tasaciones, peritaciones, estudios, informes, planificación de tareas y otros trabajos análogos de informática, de acuerdo con los conocimientos adquiridos para la tecnología específica de Ingeniería del Software.
- **CG11**: Capacidad para analizar y valorar el impacto social y medioambiental de las soluciones técnicas, comprendiendo la responsabilidad ética y profesional de la actividad del Ingeniero Técnico en Informática.
- **CG12**: Conocimiento y aplicación de elementos básicos de economía y de gestión de recursos humanos, organización y planificación de proyectos, así como la legislación, regulación y normalización en el ámbito de los proyectos informáticos.

Competencias Ingeniería del Software:

- **CIS01**: Capacidad para desarrollar, mantener y evaluar servicios y sistemas software que satisfagan todos los requisitos del usuario y se comporten de forma fiable y eficiente, sean asequibles de desarrollar y mantener y cumplan normas de calidad, aplicando las teorías, principios, métodos y prácticas de la Ingeniería del Software.
- **CIS02**: Capacidad para valorar las necesidades del cliente y especificar los requisitos software para satisfacer estas necesidades, reconciliando objetivos en conflicto mediante la búsqueda de compromisos aceptables dentro de las limitaciones derivadas del coste, del tiempo, de la existencia de sistemas ya desarrollados y de las propias organizaciones.
- **CIS03**: Capacidad de dar solución a problemas de integración en función de las estrategias, estándares y tecnologías disponibles.

- **CIS04**: Capacidad de identificar y analizar problemas y diseñar, desarrollar, implementar, verificar y documentar soluciones software sobre la base de un conocimiento adecuado de las teorías, modelos y técnicas actuales.
- **CIS05**: Capacidad de identificar, evaluar y gestionar los riesgos potenciales asociados que pudieran presentarse.
- **CIS06**: Capacidad para diseñar soluciones apropiadas en uno o más dominios de aplicación utilizando métodos de la ingeniería del software que integren aspectos éticos, sociales, legales y económicos.

Contenidos

Breve descripción del contenido

Bases de datos (BD) *vs* Recuperación de Información (RI). Modelos de RI. Evaluación y *ranking*. Consultas y operaciones textuales. Métodos de indexación específicos para RI. RI estructurada y multimedia. *Clasification* & *Clustering* en RI.

Temario de la asignatura

1. Introducción a los Sistemas de Recuperación de la Información. Fundamentos.

1.1 Qué es Recuperación de Información (RI) o *Information Retrieval* (*IR*).
1.2 Información *vs* Recuperación de datos.
1.3 Sistemas de Recuperación de la Información (SRI) *vs* Sistemas de Bases de Datos (SBD).
1.4 Arquitectura de un SRI.
1.5 Caracterización y taxonomía de modelos de RI.
1.6 SRI Web.
1.7 Motores de búsqueda.
1.8 Adquisición y transformación de textos.
1.9 Creación de índices.
1.10 Interfaces de usuario destinadas a búsquedas, Visualización.
1.11 Modelado, clasificación de los resultados mediante *ranking*.
1.12 Evaluación.

2. Documentos y Consultas. Procesamiento y análisis de textos.

2.1 Procesamiento y estimación textual.
2.2 Ley de Zipf.
2.3 Crecimiento del vocabulario.
2.4 Análisis de documentos. *Tokenization*. *Stopping*. *Stemming*.
2.5 Estructura de los documentos.
2.6 Frases. Marcado. N-gramas.
2.7 Análisis de enlaces. Texto de anclaje.
2.8 *PageRank*.
2.9 Calidad de los enlaces.
2.10 Enlaces de retorno.
2.11 Extracción de información: modelos ocultos de Markov.
2.12 Internacionalización.

Laboratorio: Análisis de documentos en un entorno personal.

3. Rastreo y recuperación Web.

3.1 Conceptos básicos. Historia del rastreo web.
3.2 Recuperación de páginas Web. Rastreadores (*web crawlers*). Algoritmo.
3.3 Arquitectura de los rastreadores web.
3.4 Control y errores de rastreo.

3.5 Frescura y edad de los documentos.
3.6 Rastreo enfocado o dirigido.
3.7 Web profunda (*Deep web*).
3.8 Mapas de sitios Web (*Sitemaps*).
3.9 Rastreo distribuido.
3.10 Rastreadores personales o de escritorio (*Desktop crawls*).
3.11 Rastreo de documentos y correos electrónicos.
3.12 Alimentación de documentos. Sindicación RSS.
3.13 Conversión de formatos estructurados y semiestructurados. Apache Tika.
3.14 Codificación. Glifos. ASCII y UTF en Unicode.
3.15 Almacenamiento documental.
3.16 Compresión y grandes ficheros. Actualizaciones.
3.17 *Google BigTable*.
3.18 Detección de duplicados y cuasi duplicados.
3.19 *Fingerprints* y método *Simhash* de búsqueda por similitud.
3.20 Eliminación de ruido y búsqueda de bloques de contenido.
Laboratorio: Desarrollo de un rastreador (*crawler*) en un entorno personal.

4. Mecanismos fundamentales de indexación.
4.1 Utilización de índices para recuperación y ranking.
4.2 Índices y ficheros invertidos (*Inverted index / files*).
4.3 Archivos de firma (*Signature files*).
4.4 Árboles de sufijos (*Suffix trees*) y vectores de sufijos (*Suffix arrays*).
4.5 Indexación multidimensional (*Multidimensional Indexing*). *R-tree*, *R⁺-tree* y *R*-tree*. *Quadtree*, *Octree* y *Kd-tree*. *Grid-files*.
4.6 Particionamiento de los datos.
4.7 Compresión.
4.8 Construcción de índices simple y mediante mezcla.
Laboratorio: Alimentación indexada de documentos al *crawler*.

5. Almacenamiento distribuido y paralelización en los SRI.
5.1 Aceleración de la RI mediante el uso del paralelismo, clústeres y distribución.
5.2 Ejecución distribuida *MapReduce*. *Hadoop*.
5.3 Almacenamiento en bases de datos *NoSQL* (*Aerospike*, *ArangoDB*, *ArcadeDB*, *Cassandra*, *CouchBase* y *CouchDB*, *DynamoDB* y *AWS*, *Elastic*, *Firestore*, HBase, MongoDB, *Neo4J*, *Redis*, *RethinkDB*, *Riak* y otras).
5.4 Persistencia políglota.
Laboratorio: Estudio completo de dos BD *NoSQL* por grupos de dos estudiantes y desarrollo de una aplicación dedicada. *Pipeline* distribuido con *Hadoop*.

6. Modelos de recuperación. Evaluación, consultas y ranking extendidos.
6.1 Recuperación booleana.
6.2 Modelo de espacio vectorial.
6.3 Modelos probabilísticos: RI como clasificación; algoritmo BM25.
6.4 *Ranking* basado en modelos de lenguaje: clasificación según probabilidad de la consulta, modelos de relevancia y retroalimentación por pseudorelevancia.
6.5 Modelo de redes de inferencia.
6.6 Búsqueda Web.
6.7 RI y aprendizaje máquina.
6.8 Modelos basados en la aplicación.
Laboratorio: propuesta y desarrollo de un *ranking* al *crawler* implementado.

7. Recuperación estructurada y multimedia.
7.1 Modelos de recuperación basados en características.

7.2 Modelos de dependencia de términos.
7.3 Recuperación estructurada.
7.4 Búsqueda experta. Sistemas *QA* de respuesta a preguntas.
7.5 Otros medios de información.
7.6 Texto ruidoso.
7.7 Transcripción hablada.
7.8 Imágenes. Vídeo. Música.
7.9 Librerías digitales.
Laboratorio: Utilización de *Apache Tika* para convertir y analizar documentos.

8. Clasificación y Clustering en la RI.

8.1 Clasificación y *Clustering*.
8.2 Clasificación por categorización. Ontologías.
8.3 Clasificadores Naïve Bayes, Perceptrones y Máquinas de vectores de soporte (SVM).
8.4 Evaluación de la clasificación. Precisión.
8.5 Clasificación y selección de características.
8.6 Medidas de distancia. Distancia editada.
8.7 *Spam*, opinión y publicidad online.
8.8 *Clustering* aglomerativo. Métodos *K-means*, *K-medoids*, *KNN* (K vecinos próximos).
8.9 *Clustering* jerárquico. Dendogramas. Métodos aglomerativos y divisivos. *Agnes* y *Diana*.
8.10 *Clustering* basado en densidad.
8.11 Evaluación del *Clustering*. Costes.
8.12 *Clustering* y búsqueda en RI.
8.13 Redes Neuronales (*Neural Networks*, NN) y Modelos de Lenguaje de Gran Tamaño (*Large Language Models*, LLM).

Laboratorio: Desarrollo opcional de un clasificador o clustering en la aplicación con BD *NoSQL* realizada (grupos de dos estudiantes).

Actividades formativas

Horas de trabajo del alumno por tema		Horas teóricas	Actividades prácticas				Actividad de seguimiento	No presencial
Tema	**Total**	**GG**	**PCH**	**LAB**	**ORD**	**SEM**	**TP**	**EP**
1	7	2		4			1	0
2	14	2		4			0	8
3	23	6		6			1	10
4	14	2		2			0	10
5	55	4		10			1	40
6	14	2		2			0	10
7	11	2		2			0	7
8	8	6		0			0	2
Evaluación **	4	4						
TOTAL	150	30		30			3	87

GG: Grupo Grande (85 estudiantes).
PCH: prácticas clínicas hospitalarias (7 estudiantes)
LAB: prácticas laboratorio o campo (15 estudiantes)
ORD: prácticas sala ordenador o laboratorio de idiomas (20 estudiantes)
SEM: clases problemas o seminarios o casos prácticos (40 estudiantes).

** Indicar el número total de horas de evaluación de esta asignatura.

TP: Tutorías Programadas (seguimiento docente, tipo tutorías ECTS).
EP: Estudio personal, trabajos individuales o en grupo, y lectura de bibliografía.

Metodologías docentes

Clases teórico-prácticas en el aula:
Clases expositivas para el desarrollo de los contenidos fundamentales de las materias. Se incluyen actividades breves, individuales o en grupo que permitan aplicar los conceptos expuestos y resolver problemas, facilitando la participación activa de los estudiantes.

Sesiones de laboratorio:
Actividades prácticas, sesiones de laboratorio guiadas, seminarios de resolución de problemas, etc., en grupos bajo la dirección del profesor. Se podrán incluir actividades previas y posteriores a las sesiones de laboratorio y seminario que ayuden a conseguir los objetivos propuestos.

Realización de actividades, trabajos y estudio por parte del estudiante, de manera autónoma, individualmente o en grupo:
Las actividades que el estudiante desarrollará de manera no presencial estarán orientadas principalmente a la adquisición de conocimientos básicos en el ámbito de la Informática y al desarrollo de los proyectos y trabajos solicitados, bien individualmente o en grupo.

De manera más específica, la metodología docente se centra en:

- Clases teóricas con apoyo de material audiovisual, presentándose los contenidos básicos de cada tema. Tras una breve introducción de los contenidos que se pretenden desarrollar en la clase, y de los comentarios oportunos de los conceptos asumidos en clases anteriores, el desarrollo docente se realiza apoyado con transparencias, textos bibliográficos y artículos de interés y videos disponibles de expertos en el campo a tratar. Se motiva a los estudiantes a intervenir en cualquier momento para hacer las clases más dinámicas y facilitar el aprendizaje. Los temas terminan con la exposición de las conclusiones más relevantes de los conceptos tratados.
- Trabajos de investigación realizados por los estudiantes sobre conceptos de Recuperación de la Información, realización de un sistema de recuperación personal con un rastreador o *crawler*, un analizador documental, el almacenamiento indexado y la recuperación mediante ranking; además del modelo de programación *MapReduce* con Hadoop y Bases de datos NoSQL. Los trabajos se desarrollarán en grupos de dos estudiantes generalmente.
- Presentación oral (máximo 15 minutos por grupo) y defensa de algunos de los trabajos propuestos por el profesor por parte de todos los componentes de los grupos de trabajo.
- Tutorías (con opción a ser virtuales) para consultar cualquier duda relacionada con los contenidos, organización y planificación de la materia.
- Campus Virtual y Blogs personales de los estudiantes y de la asignatura. En el Campus Virtual se dejará todo el material utilizado y necesario para el desarrollo del curso, así como las indicaciones de los trabajos a realizar, sus cometidos y plazos de entrega en el Blog de la asignatura. Asimismo, cada estudiante irá plasmando la evolución de sus trabajos y conocimientos en un blog individual dependiente del blog principal de la asignatura. Estos blogs de los estudiantes servirán para ser evaluados y comentados por el resto de los estudiantes como un pilar fundamental de la metodología docente de esta asignatura.

Resultados de aprendizaje

Al finalizar el curso satisfactoriamente, el alumno:

- Conoce y aplica en actividades avanzadas las competencias transversales fundamentales de la profesión.
- Describe las particularidades de la Recuperación de Información y entiende las limitaciones de los sistemas de bases de datos para dar solución a los problemas que aparecen en este contexto.
- Aplica modelos de representación de información textual para realizar operaciones de almacenamiento y búsqueda en Recuperación de Información.
- Conoce las técnicas para la evaluación del rendimiento en sistemas de Recuperación de Información.
- Aplica algoritmos de *ranking* de documentos en respuestas a consultas en la Web.
- Identifica métodos de indexación específicos para Recuperación de Información.
- Aplica los métodos y técnicas de Recuperación de Información en bibliotecas digitales.
- Conoce modelos de comunicación multimedia basados en lenguajes de marcado.

Sistemas de evaluación

Según la normativa de evaluación vigente en la Universidad de Extremadura, los estudiantes podrán superar la asignatura mediante la modalidad de evaluación continua o mediante la modalidad de evaluación global.
Cada estudiante deberá decidir a qué tipo de evaluación se acoge en cada una de las convocatorias del curso y comunicarlo al profesorado durante el primer cuarto del periodo de impartición de la asignatura. Dicha comunicación se realizará mediante una consulta que estará disponible en el aula virtual. Si no se realiza esta comunicación en el plazo establecido se entenderá que el estudiante opta por la evaluación continua.

Evaluación continua:
La opción deseable para llevar a cabo plenamente la metodología docente utilizada. Se valora la asistencia presencial a las clases teórico-prácticas en el aula. En las mismas y en las sesiones de laboratorio se propondrán la realización de trabajos prácticos (individuales y en grupo) y la presentación y exposición de ciertos trabajos específicos planteados.

El trabajo de cada estudiante quedará reflejado en un *blog* docente individual ligado a la asignatura, que se irá construyendo a lo largo del curso. Paralelamente, los blogs de los estudiantes servirán para que los estudiantes coevalúen el trabajo realizado entre ellos. La evaluación final se decidirá por parte del profesor basándose en la ponderación de todos los trabajos realizados, su presentación, el blog construido y en el contenido justificado de las coevaluaciones de los estudiantes. Cada uno de estos trabajos será evaluado individualmente y la calificación será la proporcionada por la media de todas las calificaciones obtenidas.

Evaluación global:
Tanto si el profesor estima muy baja la participación por parte del estudiante a lo largo del curso, como si el estudiante decide ser evaluado sin utilizar evaluación continua durante el curso ordinario, así como para cualquier convocatoria extraordinaria, cabe la evaluación basada en el resultado de calificar el 50% de su nota mediante un examen final. El otro 50% de la nota vendrá dada por la realización y entrega de todos los

trabajos obligatorios propuestos durante el curso. El estudiante deberá realizar y entregar todos estos trabajos antes de la fecha de examen indicada en el calendario de exámenes fijado por la dirección del centro.

Sistema de revisión de exámenes y pruebas de evaluación continua:
La revisión de exámenes y pruebas se realizará de acuerdo a la normativa de evaluación vigente.
Para el examen realizado en cada convocatoria oficial, el alumno podrá comentar y revisar los resultados del mismo en las fechas y horarios que se indiquen en la publicación de las calificaciones provisionales.
Para el resto de pruebas de evaluación continua que se realicen durante el semestre, la revisión se realizará en el plazo máximo de 10 días hábiles después de la publicación de la calificación. Dependiendo del tipo de prueba, esta revisión podrá realizarse en el horario de tutorías de libre acceso del profesor o en un horario especificado de manera particular para la revisión de dicha prueba.

Recomendación importante:
Se recomienda la asistencia a todas las clases presenciales y a las tutorías, dado el alto contenido práctico de la asignatura. Sobretodo porque comprender bien todos los conceptos impartidos van a permitir realizar los ejercicios prácticos de manera más aplicada. La participación continuada en el estudio y el desarrollo de los contenidos, sin dejar para el final la materia, se considera la vía idónea para lograr los objetivos de aprendizaje.

Bibliografía (básica y complementaria)

Básica: (por orden de recomendación)

[1] Manning, C. D., Raghavan, P. & Schütze, H. (2008) "*Introduction to Information Retrieval*", 1st ed., Cambridge.
[2] Leskovec, J., Rajaraman, A. & Ullman, J. D. (2020) "*Mining of Massive Datasets*", 3th ed., Cambridge

Complementaria:

[3] Baeza-Yates, R. & Ribeiro-Neto, B. (2011) "*Modern Information Retrieval*", 2nd ed., Addison-Wesley.
[4] Croft, B., Metzler, D. & Strohman, T. (2009) "*Search Engines: Information Retrieval in Practice*", 1st ed., Pearson.
[5] White, T. (2015). "*Hadoop. The Definitive Guide*", 4th ed., O'Reilly.

Otros recursos y materiales docentes complementarios

Recursos de laboratorio y trabajo no presencial dejados o enlazados en el Campus Virtual y en el Blog principal de la asignatura.