Guía docente
270657 - IR - Recuperación de la Información

Última modificación: 12/07/2021

Unidad responsable: Facultad de Informática de Barcelona
Unidad que imparte: 723 - CS - Departamento de Ciencias de la Computación.

Titulación: MÁSTER UNIVERSITARIO EN INNOVACIÓN E INVESTIGACIÓN EN INFORMÁTICA (Plan 2012). (Asignatura optativa).

Curso: 2021 **Créditos ECTS:** 6.0 **Idiomas:** Inglés

PROFESORADO

Profesorado responsable: RAMON FERRER CANCHO

Otros: Primer quadrimestre:
RAMON FERRER CANCHO - 10

CAPACIDADES PREVIAS

Las supuestas en el ingreso en el MIRI más las proporcionadas por la fase de formación común.

COMPETENCIAS DE LA TITULACIÓN A LAS QUE CONTRIBUYE LA ASIGNATURA

Específicas:
CEC1. Capacidad para aplicar el método científico en el estudio y análisis de fenómenos y sistemas en cualquier ámbito de la Informática, así como en la concepción, diseño e implantación de soluciones informáticas innovadoras y originales.
CEC2. Capacidad para el modelado matemático, cálculo y diseño experimental en centros tecnológicos y de ingeniería de empresa, particularmente en tareas de investigación e innovación en todos los ámbitos de la Informática.
CEC3. Capacidad para aplicar soluciones innovadoras y realizar avances en el conocimiento que exploten los nuevos paradigmas de la Informática, particularmente en entornos distribuidos.

Genéricas:
CG1. Capacidad para aplicar el método científico en el estudio y análisis de fenómenos y sistemas en cualquier ámbito de la Informática, así como en la concepción, diseño e implantación de soluciones informáticas innovadoras y originales.
CG3. Capacidad para el modelado matemático, cálculo y diseño experimental en centros tecnológicos y de ingeniería de empresa, particularmente en tareas de investigación e innovación en todos los ámbitos de la Informática.
CG5. Capacidad para aplicar soluciones innovadoras y realizar avances en el conocimiento que exploten los nuevos paradigmas de la Informática, particularmente en entornos distribuidos.

Transversales:
CTR4. USO SOLVENTE DE LOS RECURSOS DE INFORMACIÓN: Gestionar la adquisición, la estructuración, el análisis y la visualización de datos e información del ámbito de la ingeniería informática y valorar de forma crítica los resultados de esta gestión.
CTR5. ACTITUD FRENTE AL TRABAJO: Tener motivación para la realización profesional y para afrontar nuevos retos, así como una visión amplia de las posibilidades de la carrera profesional en el ámbito de la Ingeniería en Informática. Tener motivación por la calidad y la mejora continua, y actuar con rigor en el desarrollo profesional. Capacidad de adaptación a los cambios organizativos o tecnológicos. Capacidad de trabajar en situaciones de falta de información y/o con restricciones temporales y/o de recursos.
CTR6. RAZONAMIENTO: Capacidad de razonamiento crítico, lógico y matemático. Capacidad para resolver problemas dentro de su área de estudio. Capacidad de abstracción: capacidad de crear y utilizar modelos que reflejen situaciones reales. Capacidad de diseñar y realizar experimentos sencillos, y analizar e interpretar sus resultados. Capacidad de análisis, síntesis y evaluación.

Básicas:
CB6. Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.
CB7. Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios
CB8. Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades.
CB9. Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

METODOLOGÍAS DOCENTES

Sesiones de teoría + problemas de 3 horas semanales. Las 2 primeras horas de cada sesión son de tipo teórico, y la tercera se dedica a problemas. Para cada sesión, el estudiante tendrá que entregar algunos problemas propuestos pero no resueltos en el anterior.

Sesiones de laboratorio de 1 hora semanal. Para muhcas de las sesiones, el estudiante tendrá que entregar un informe del trabajo hecho y resultados obtenidos en alrededor de 2 semanas.

El funcionamiento de cada tipo de sesión se describe en el apatado "Actividades".

Además, hacia final de curso cada estudiante presentará ante los profesores y otros matriculados un artículo científico relacionado con la temática de la asignatura, como si se tratara de una presentación en congreso. Hacia la semana 8 se publicará una lista de posibles artículos de los cuales cada estudiante elegirá uno. Alternativamente, podrá proponer un artículo de su elección para que la aprueben los profesores. El día de las presentaciones se anunciará con al menos 2 meses de tiempo y el orden y hora exactos de las presentaciones con al menos 1 semana de antelación.

OBJETIVOS DE APRENDIZAJE DE LA ASIGNATURA

HORAS TOTALES DE DEDICACIÓN DEL ESTUDIANTADO

Tipo	Horas	Porcentaje
Horas grupo grande	54,0	36.00
Horas aprendizaje autónomo	96,0	64.00

Dedicación total: 150 h

CONTENIDOS

Introducción

Descripción:
Necesidad de técnicas de búsqueda y análisis de información masiva. Búsqueda y análisis vs. bases de datos. Proceso de recuperación de la información. Preproceso y análisis léxico.

Modelos de recuperación de la información

Descripción:
Definición formal y conceptos básicos: Modelos abstractos de documentos y lenguajes de interrogación. Modelo booleano. Modelo vectorial. Latent Semantic Indexing.

Implementación: Indexación y búsquedas

Descripción:
Ficheros inversos y ficheros de firmas. Compresión de índices. Ejemplo: Implementación eficiente de la regla del coseno con medida tf-idf. Ejemplo: Lucene.

Evaluación en recuperación de la información

Descripción:
Recall y precisión. Otras medidas de rendimiento. Colecciones de referencia. "Relevance feedback" y "query expansion".

Búsqueda en internet

Descripción:
Ranking y relevancia para modelos web. Algoritmo PageRank. Crawling. Arquitectura de un sistema simple de búsqueda en la web.

Arquitectura de sistemas para la gestión de información masiva

Descripción:
Escalabilidad, alto rendimento y tolerancia a fallos: el caso de buscadores web masivos. Arquitecturas distribuidas. Ejemplo: Hadoop.

Análisis de redes

Descripción:
Parámetros descriptivos y características de las redes: grado, diámetro, redes "small-world", entre otros. Algoritmos sobre redes: clustering, detección de comunidades y de nodos influyentes, reputación, entre otros.

Sistemas de información basados en análisis de información masiva. Combination with other technologies.

Descripción:
"Search Engine Optimization". Uso de técnicas de recuperación de la información en combinación con Minería de Datos y Aprendizaje. Sistemas de recomendación.

ACTIVIDADES

Desarrollo teórico de los temas 1 a 8 del curso

Descripción:
El alumno atenderá a la exposición del profesor y participará activamente en la discusión inicial del reto a resolver en la sesión.

Dedicación: 52h
Grupo grande/Teoría: 26h
Aprendizaje autónomo: 26h

Ejercicios sobre los temas 1 a 8 de la asignatura

Descripción:
En cada sesión, el profesor plantea una colección de problemas (orientativamente, entre 4 y 7) del tema que acaba de tratarse teòricamente. A continuación se resuelven conjuntamente algunos de los problemas propuestos (orientativamente, 3). Los estudiantes han de resolver el resto de los problemas fuera de horas de clase, y entregarlos al inicio de la siguiente sesión. Parte de la sesión se dedica a comentar las dudas que puedan haber salido en la resolución de estos problemas pendientes de la sesión anterior.

Dedicación: 39h
Grupo mediano/Prácticas: 13h
Aprendizaje autónomo: 26h

Trabajo de laboratorio sobre los temas 1 a 8

Descripción:
El profesor planteará un trabajo de tipo práctico relacionado con los temas vistos más recientemente. Éste puede consistir en el análisis de unos datos proporcionados (o que haga falta buscar), implementar uno de los algoritmos vistos en clase, o proporcionar una solución a un caso concreto de necesidad de técnicas de recuperación de la información. El estudiante completa el trabajo en la medida de lo posible dentro de las horas de clase, aunque algún tiempo fuera de clase pueden ser necesario. En muchas de las sesiones será necesario entregar un informe del trabajo hecho y resultados obtenidos, a entregar en el plazo que se determinará en cada caso (orientativamente, 2 semanas).

Dedicación: 26h
Grupo pequeño/Laboratorio: 13h
Aprendizaje autónomo: 13h

Examen final

Descripción:
Examen final sobre el contenido de toda la asignatura

Dedicación: 18h
Actividades dirigidas: 3h
Aprendizaje autónomo: 15h

Estudio y presentación de un artículo científico

Descripción:
Estudio y presentación de un artículo científico relacionado con la asignatura

Dedicación: 13h
Actividades dirigidas: 3h
Aprendizaje autónomo: 10h

SISTEMA DE CALIFICACIÓN

Sean:

- NF la nota del examen final, de 0 a 10,
- NE la nota de las entregas de ejercicios, de 0 a 10,
- NL la nota de las prácticas de laboratorio, de 0 a 10,
- NA la nota de la presentación de un artículo científico,

todas en el rango de 0 a 10.

La nota final de la assignatura es 0.3*NF + 0.25*NL + 0.25*NE + 0.2*NA.

BIBLIOGRAFÍA

Básica:

- Manning, C.D.; Raghavan, P.; Schütze, H. Introduction to information retrieval. Cambridge University Press, 2008. ISBN 9780521865715.
- Croft, W.B.; Metzler, D.; Strohman, T. Search engines: information retrieval in practice. Pearson, 2010. ISBN 9780131364899.
- Russell, M.A.; Klassen, M. Mining the social web: data mining Facebook, Twitter, LinkedIn, Instagram, Github, and more. 3rd ed. O'Reilly Media, 2018. ISBN 9781491973509.
- McCandless, M.; Hatcher, E.; Gospodnetic, O. Lucene in action. 2nd ed. Manning, 2010. ISBN 9781933988177.
- Baeza-Yates, R.; Ribeiro-Neto, B. Modern information retrieval: the concepts and technology behind search. 2nd ed. Addison-Wesley / Pearson, 2011. ISBN 9780321416919.

RECURSOS

Enlace web:

- http://www.cs.upc.edu/~IR-MIRI/