

Objetivos de la asignatura

El gran volumen de datos generados diariamente de tipo no estructurado o semiestructurado hace necesaria la aplicación de técnicas de rastreo, procesamiento, indexación y recuperación que proporcionen acceso eficiente y eficaz a ese tipo de contenidos, que pueden estar en una gran variedad de formatos (por ejemplo, texto, web, xml, imágenes, video).

Esta materia proporciona formación en tecnología de recuperación de información, necesaria para la gestión de masivos repositorios documentales, así como aspectos relacionados como el procesamiento de lenguaje natural y la extracción de información. También se tratan otras tecnologías de Minería de Textos y Web como la clasificación y el agrupamiento documental.

Contenidos

Motores de búsqueda y Recuperación de Información

Procesamiento de Texto, construcción de índices invertidos y compresión

Modelos de recuperación de información (vectoriales, probabilísticos), realimentación de relevancia y expansión de consultas.

Recuperación de Información semi-estructurada: recuperación XML.

Evaluación de sistemas de RI

Búsqueda web (rastreo, indexación, análisis de enlaces)

Recuperación de Información multimedia (imágenes, video)

Extracción de información

Clasificación y agrupamiento documental

Otros aspectos de Minería de Texto y Minería Web

Bibliografía básica y complementaria

Textos basicos:

Zhai, C., Massung, S. Text Data Management and Analysis: a Practical Introduction to Information Retrieval and Text Mining. Association for Computing Machinery and Morgan & Claypool, New York, NY, USA. 2016.

Textos complementarios:

Croft, B, Metzler, D, Strohman, T.. Search Engines: Information Retrieval in practice, Pearson 2010.

Cacheda, F., Fernández-Luna, J.M., Huete, J. (eds) Recuperación de Información. Un enfoque práctico y multidisciplinar. Madrid: Ra-Ma, 2011

Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, Introduction to Information Retrieval, Cambridge University Press. 2008

Baeza-Yates, R., Ribeiro-Neto, B. Modern Information Retrieval: the concepts and technology behind search, 2nd edition.

Gospodnetic, O e Hatcher, E. "Lucene in Action", Manning, 2005

Competencias

El alumno conocerá distintos modelos de recuperación de información y análisis de textos y comprenderá sus beneficios e inconvenientes para el tratamiento de masivos repositorios de datos de distinta índole.

El alumno será capaz de implementar software para el rastreo, procesamiento, indexación y recuperación de datos no estructurados o semi-estructurados

El alumno comprenderá la metodología de evaluación de sistemas de búsqueda y será capaz de validar empíricamente estrategias de búsqueda y clasificación de contenidos

El alumno conocerá las principales estrategias de Minería de Texto y Web y podrá implementarlas en proyectos reales

El alumno adquirirá la habilidad necesaria para la búsqueda, selección y manejo de recursos (bibliografía, software, etc.) relacionados con esta materia

Competencias de la titulación que se trabajan (ver memoria título):

- Básicas: CB6, CB7, CB8, CB10
- Transversales/Generales: G1, G2, G4, T4, T7
- Específicas: E1, E2

Metodología de enseñanza

Clases teóricas, en las que se expone el contenido de cada tema. El alumno dispondrá de copias de las transparencias con anterioridad y el profesor promoverá una actitud activa, realizando preguntas que permitan aclarar aspectos concretos y dejando cuestiones abiertas para la reflexión del alumno.

Clases prácticas con uso de ordenador, que permiten al alumno familiarizarse desde un punto de vista práctico con las cuestiones expuestas en las clases teóricas.

Actividades formativas de carácter presencial y su relación con las competencias de la titulación:

Clases teóricas: impartidas por el profesor y exposición de seminarios: CB6, CB8, CB10, E1, E2

Clases prácticas de laboratorio, resolución de problemas y casos prácticos: CB7, CB8, CB10, E1, E2

Tutorías programadas: orientación para la realización de los trabajos individuales o en grupo, resolución de dudas y actividades de evaluación continua: CB6, CB7, CB8, CB10, G1, T4, G2, G4, T5, E1, E2

Examen: CB6, CB7, CB8, CB10, G1, T4, G2, G4, T5, E1, E2

Actividades formativas de carácter no presencial y su relación con las competencias de la titulación:

Trabajo personal del alumno: consulta de bibliografía, estudio autónomo, desarrollo de actividades programadas, preparación de presentaciones y trabajos : E1, E2, CB6, CB7, CB8, CB10, G1, T4, G2, G4, T5

Sistema de evaluación

Realización de prácticas: 60%

Pruebas periódicas y/o examen final: 30%

Seguimiento continuado y objetivable de una participación activa: 10%

Para superar la materia es necesario que al menos un 10% de la nota total provenga de las Pruebas periódicas y/o examen final

La evaluación en la segunda oportunidad es exactamente igual que en la primera oportunidad (entrega de prácticas, realización de pruebas y seguimiento de participación activa).

Tiempo de estudio y trabajo personal

Clases de pizarra: 18 h presenciales + 33 h trabajo autonomo del alumno

Clases de prácticas: 13.5 h presenciales + 33 h trabajo autonomo del alumno

Tutorías y Actividades de evaluacion: 4 h presenciales + 11 h trabajo autonomo del alumno

Total: 112.5 h