Guía docente
340455 - REIN-I7P23 - Recuperación de la Información

Última modificación: 30/06/2023

Unidad responsable: Escuela Politécnica Superior de Ingeniería de Vilanova i la Geltrú
Unidad que imparte: 723 - CS - Departamento de Ciencias de la Computación.

Titulación: GRADO EN INGENIERÍA INFORMÁTICA (Plan 2018). (Asignatura optativa).

Curso: 2023 **Créditos ECTS:** 6.0 **Idiomas:** Catalán

PROFESORADO

Profesorado responsable: Neus Català i Roig

Otros: Neus Català i Roig

CAPACIDADES PREVIAS

- Saber aplicar los conceptos básicos de álgebra lineal, matemática discreta, probabilidad y estadística.
- Saber programar en lenguajes basados en objetos, incluyendo herencia entre clases.
- Conocer las principales estructuras de datos para el acceso eficiente a información y sus implementaciones (listas, hashing, árboles, grafos, heaps). Ser capaz de utilizarlas para construir programas eficientes. Poder analizar el tiempo de ejecución y memoria usada por un programa de dificultad media. Tener una cierta idea de la diferencia en tiempo de acceso entre memoria principal y memoria secundaria.
- Conocer los elementos principales de una base de datos relacional y de lenguajes de acceso tipo SQL.

REQUISITOS

Haber aprobado ESTA, AMEP y DABD o al menos estar matriculado/a.

COMPETENCIAS DE LA TITULACIÓN A LAS QUE CONTRIBUYE LA ASIGNATURA

Específicas:
1. CECO7. Capacidad para conocer y desarrollar técnicas de aprendizaje computacional y diseñar e implementar aplicaciones y sistemas que las utilicen, incluyendo las dedicadas a extracción automática de información y conocimiento a partir de grandes volúmenes de datos.
4. CEIS6. 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.
3. CEIS4. 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.
2. CEIS1. 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.

Transversales:
5. EMPRENDEDURÍA E INNOVACIÓN: Conocer y entender la organización de una empresa y las ciencias que definen su actividad; capacidad para entender las normas laborales y las relaciones entre la planificación, las estrategias industriales y comerciales, la calidad y el beneficio.

METODOLOGÍAS DOCENTES

La asignatura consta de:
- 2 horas a la semana de clases presenciales en el aula en las que el profesor expone los contenidos (teoría y problemas),
- 2 horas a la semana en el aula de laboratorio en las que se aplican los conceptos y métodos aprendidos mediante la resolución de problemas haciendo uso de herramientas específicas del àrea de la Recuperación de la Información (Information Retrieval).

OBJETIVOS DE APRENDIZAJE DE LA ASIGNATURA

La cantidad de información almacenada digitalmente en organizaciones, o colectivamente en la web, es hoy en día suficientemente voluminosa como para que encontrar aquello que se busca sea generalmente complicado. El campo conocido como "Information Retrieval" contempla métodos para organizar la información de forma que sea posible después para los usuarios encontrar la información de forma cómoda y eficiente.

La asignatura cubre las técnicas básicas de búsqueda de documentación textual basada en palabras clave. También examina el caso de la búsqueda en la web, donde la presencia de hiperenlaces puede usarse no sólo para dirigir la búsqueda sino para valorar el interés de cada página -es el caso del conocido algoritmo PageRank utilizado por Google. Se ve la extensión de este tipo de técnicas a los casos de búsqueda personalizada y sistemas de recomendación. Finalmente, se incluye una breve introducción a la búsqueda semántica y a los modelos neuronales de recuperación de la información.

HORAS TOTALES DE DEDICACIÓN DEL ESTUDIANTADO

Tipo	Horas	Porcentaje
Horas grupo grande	30,0	20.00
Horas grupo pequeño	30,0	20.00
Horas aprendizaje autónomo	90,0	60.00

Dedicación total: 150 h

CONTENIDOS

1. 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.

Actividades vinculadas:
Actividad 1: Control 1
Actividad 3: Laboratorio

Dedicación: 11h
Grupo grande/Teoría: 1h 30m
Grupo pequeño/Laboratorio: 2h 30m
Aprendizaje autónomo: 7h

2. Modelos de recuperación de la información

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

Actividades vinculadas:
Actividad 1: Control 1
Actividad 3: Laboratorio

Dedicación: 12h
Grupo grande/Teoría: 1h 30m
Grupo pequeño/Laboratorio: 3h 30m
Aprendizaje autónomo: 7h

3. 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: ElasticSearch.

Actividades vinculadas:
Actividad 1: Control 1
Actividad 3: Laboratorio

Dedicación: 10h
Grupo grande/Teoría: 0h 30m
Grupo pequeño/Laboratorio: 2h 30m
Aprendizaje autónomo: 7h

4. Evaluación en recuperación de la información

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

Actividades vinculadas:
Actividad 1: Control 1
Actividad 3: Laboratorio

Dedicación: 10h
Grupo grande/Teoría: 0h 30m
Grupo pequeño/Laboratorio: 2h 30m
Aprendizaje autónomo: 7h

5. Búsqueda en internet

Descripción:
Ranking y relevancia para modelos web. Algoritmos PageRank y HITS. Crawling. Arquitectura de un sistema simple de búsqueda en internet.

Actividades vinculadas:
Actividad 2: Control 2
Actividad 3: Laboratorio

Dedicación: 16h
Grupo grande/Teoría: 3h
Grupo pequeño/Laboratorio: 6h
Aprendizaje autónomo: 7h

6. Arquitectura de sistemas para la gestión de información masiva

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

Actividades vinculadas:
Actividad 2: Control 2
Actividad 3: Laboratorio

Dedicación: 12h 30m
Grupo grande/Teoría: 3h
Grupo pequeño/Laboratorio: 6h
Aprendizaje autónomo: 3h 30m

7. Sistemas de información basados en análisis de información masiva

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

Actividades vinculadas:
Actividad 2: Control 2
Actividad 3: Laboratorio

Dedicación: 10h 30m
Grupo grande/Teoría: 2h
Grupo pequeño/Laboratorio: 5h
Aprendizaje autónomo: 3h 30m

8. Búsqueda semántica y modelos neuronales de recuperación de la información

Descripción:
Búsqueda semántica: cómo conseguir que las búsquedas incluyan significados y contextos, no solo palabras clave. Word embeddings y sentence embeddings. Modelos neuronales de recuperación de la información. Re-ranking.

Actividades vinculadas:
Actividad 2: Control 2
Actividad 3: Laboratorio

Dedicación: 15h
Grupo grande/Teoría: 2h
Grupo pequeño/Laboratorio: 6h
Aprendizaje autónomo: 7h

SISTEMA DE CALIFICACIÓN

La asignatura contiene los siguientes actos de evaluación:
- Informes de las actividades y/o prácticas de las sesiones de laboratorio (L).
- Control parcial 1 realizado en el periodo de evaluación parcial (C1).
- Control parcial 2 realizado en el periodo de evaluación final (C2).

La ponderación de las calificaciones obtenidas es la siguiente:

0.4*L + 0.3*C1 + 0.3*C2

NORMAS PARA LA REALIZACIÓN DE LAS PRUEBAS.

Los informes de las actividades de laboratorio se entregan de forma no presencial en el plazo indicado para cada sesión.

Los controles parciales son presenciales e individuales.

BIBLIOGRAFÍA

Básica:
- Russell, Matthew A; Klassen, Mikhail. Mining the social web : data mining Facebook, Twitter, LinkedIn, Instagram, Github, and more [en línea]. 3rd ed. Sebastopol, [California]: O'Reilly Media, 2018 [Consulta: 14/02/2024]. Disponible a: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=5611114. ISBN 9781491973509.
- Baeza-Yates, Ricardo ; Ribeiro-Neto, Berthier. Modern information retrieval : the concepts and technology behind search. 2nd ed. Harlow [etc.]: Addison-Wesley, 2011. ISBN 9780321416919.
- Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich. Introduction to information retrieval [en línea]. New York: Cambridge University Press, 2008 [Consulta: 25/03/2022]. Disponible a: https://nlp.stanford.edu/IR-book/html/htmledition/irbook.html. ISBN 9780521865715.
- Croft, W. Bruce; Metzler, Donald; Strohman, Trevor. Search engines : information retrieval in practice. Boston [etc.]: Pearson, 2010. ISBN 9780131364899.

RECURSOS

Otros recursos:
Enlaces web:

- An Introduction to Neural Information Retrieval, by Bhaskar Mitra and Nick Craswell (https://arxiv.org/abs/1705.01509)

- The Anatomy of a Large-Scale Hypertextual Web Search Engine, by Sergey Brin and Lawrence Page (http://infolab.stanford.edu/backrub/google)

- The Heart of the Elastic Stack (https://www.elastic.co/products/elasticsearch)