

Guia docent

340455 - REIN-I7P23 - Recuperació de la Informació

Última modificació: 03/04/2024

Unitat responsable: Escola Politècnica Superior d'Enginyeria de Vilanova i la Geltrú

Unitat que imparteix: 723 - CS - Departament de Ciències de la Computació.

Titulació: GRAU EN ENGINYERIA INFORMÀTICA (Pla 2018). (Assignatura optativa).

Curs: 2024

Crèdits ECTS: 6.0

Idiomes: Català

PROFESSORAT

Professorat responsable: Neus Català i Roig

Altres: Neus Català i Roig

CAPACITATS PRÈVIES

- Saber aplicar els conceptes bàsics d'àlgebra lineal, matemàtica discreta, probabilitat i estadística.
- Saber programar en llenguatges basats en objectes, incloent l'herència entre classes.
- Conèixer les principals estructures de dades per a l'accés eficient a informació i les seves implementacions (l·listes, hashing, arbres, grafs, heaps). Ser capaç d'usar-les per a construir programes eficients. Poder analitzar el temps d'execució i memòria usat per un programa de dificultat mitjana. Tenir una certa idea de la diferència en temps d'accés entre memòria principal i memòria secundària.
- Conèixer els elements principals d'una base de dades relacional i de llenguatges d'accés tipus SQL.

REQUISITS

Haver aprovat ESTA, AMEP i DABD o almenys estar-hi matriculat/da.

COMPETÈNCIES DE LA TITULACIÓ A LES QUALS CONTRIBUEIX L'ASSIGNATURA

Específiques:

1. CECO7. Capacitat per a conèixer i desenvolupar tècniques d'aprenentatge computacional i dissenyar i implementar aplicacions i sistemes que les utilitzin, incloent les dedicades a extracció automàtica d'informació i coneixement a partir de grans volums de dades.
4. CEIS6. Capacitat per a dissenyar solucions apropiades en un o més dominis d'aplicació utilitzant mètodes de l'enginyeria del programari que integrin aspectes ètics, socials, legals i econòmics.
3. CEIS4. Capacitat d'identificar i analitzar problemes i dissenyar, desenvolupar, implementar, verificar i documentar solucions programari sobre la base d'un coneixement adequat de les teories, models i tècniques actuals.
2. CEIS1. Capacitat per a desenvolupar, mantenir i avaluar serveis i sistemes programari que satisfacin tots els requisits de l'usuari i es comportin de forma fiable i eficient, siguin assequibles de desenvolupar i mantenir i compleixin normes de qualitat, aplicant les teories, principis, mètodes i pràctiques de l'Enginyeria del Programari.

Transversals:

5. EMPRENEDORIA I INNOVACIÓ: Conèixer i comprendre l'organització d'una empresa i les ciències que regeixen la seva activitat; capacitat per comprendre les regles laborals i les relacions entre la planificació, les estratègies industrials i comercials, la qualitat i el benefici.

METODOLOGIES DOCENTS

L'assignatura consta de:

- 2 hores a la setmana de classes presencials a l'aula en les que el professor exposa els continguts (teoria i problemes),
- 2 hores a la setmana a l'aula de laboratori en les que s'apliquen els conceptes i mètodes apresos mitjançant la resolució de problemes fent ús d'eines específiques de l'àrea de la Recuperació de la Informació (Information Retrieval).

OBJECTIUS D'APRENENTATGE DE L'ASSIGNATURA

La quantitat d'informació emmagatzemada digitalment en moltes organitzacions, o col·lectivament en la web, és avui dia prou gran perquè trobar allò que es busca sigui generalment complicat. El camp conegut com a "Information Retrieval" tracta dels mètodes per organitzar informació i permetre després als usuaris trobar-la de manera còmoda i eficient.

L'assignatura cobreix les tècniques bàsiques de cerca de documentació textual basada en paraules clau. També examina el cas de la cerca en la web, on la presència d'hiperenllaços pot usar-se no sols per dirigir la cerca sinó per valorar l'interès de cada pàgina -és el cas del conegut algorisme PageRank utilitzat per Google. Es veu l'extensió d'aquestes tècniques als casos de cerca personalitzada i sistemes de recomanació. Finalment, es fa una breu introducció a la cerca semàntica i als models neuronals de recuperació de la informació.

HORES TOTALES DE DEDICACIÓ DE L'ESTUDIANTAT

Tipus	Hores	Percentatge
Hores grup petit	30,0	20.00
Hores aprenentatge autònom	90,0	60.00
Hores grup gran	30,0	20.00

Dedicació total: 150 h

CONTINGUTS

1. Introducció

Descripció:

Necessitat de les tècniques de cerca i anàlisi d'informació massiva. Cerca i anàlisi vs. bases de dades. Procés de recuperació de la informació. Preprocés i anàlisi lèxica.

Activitats vinculades:

Activitat 1: Control 1

Activitat 3: Laboratori

Dedicació: 11h

Grup gran/Teoria: 1h 30m

Grup petit/Laboratori: 2h 30m

Aprenentatge autònom: 7h

2. Models de recuperació de la informació

Descripció:

Definició formal i conceptes bàsics: Models abstractes de documents i llenguatges d'interrogació. Model booleà. Model vectorial.

Activitats vinculades:

Activitat 1: Control 1

Activitat 3: Laboratori

Dedicació: 12h

Grup gran/Teoria: 1h 30m

Grup petit/Laboratori: 3h 30m

Aprenentatge autònom: 7h

3. Implementació: Indexació i cerques

Descripció:

Fitxers invertits i fitxers de signatures. Compressió d'índexos. Exemple: Implementació eficient de la regla del cosinus amb mesura tf-idf. Exemple: Elasticsearch.

Activitats vinculades:

Activitat 1: Control 1

Activitat 3: Laboratori

Dedicació: 10h

Grup gran/Teoria: 0h 30m

Grup petit/Laboratori: 2h 30m

Aprenentatge autònom: 7h

4. Avaluació en recuperació de la informació

Descripció:

Recall i precisió. Altres mesures de rendiment. Col·leccions de referència. "Relevance feedback" i "query expansion".

Activitats vinculades:

Activitat 1: Control 1

Activitat 3: Laboratori

Dedicació: 10h

Grup gran/Teoria: 0h 30m

Grup petit/Laboratori: 2h 30m

Aprenentatge autònom: 7h

5. Cerca a internet

Descripció:

Rànquing i rellevància per a models web. Algorismes PageRank i HITS. Crawling. Arquitectura d'un sistema simple de cerca a la web.

Activitats vinculades:

Activitat 2: Control 2

Activitat 3: Laboratori

Dedicació: 16h

Grup gran/Teoria: 3h

Grup petit/Laboratori: 6h

Aprenentatge autònom: 7h

6. Arquitectura de sistemes per a la gestió d'informació massiva

Descripció:

Escalabilitat, alt rendiment i tolerància a fallides: el cas de cercadors web massius. Arquitectures distribuïdes. Exemple: Hadoop.

Activitats vinculades:

Activitat 2: Control 2

Activitat 3: Laboratori

Dedicació: 12h 30m

Grup gran/Teoria: 3h

Grup petit/Laboratori: 6h

Aprenentatge autònom: 3h 30m

7. Sistemes d'informació basats en l'explotació d'informació massiva

Descripció:

"Search Engine Optimization". Utilització de tècniques de recuperació de la informació en combinació amb Mineria de Dades i Aprenentatge. Sistemes de recomanació.

Activitats vinculades:

Activitat 2: Control 2

Activitat 3: Laboratori

Dedicació: 10h 30m

Grup gran/Teoria: 2h

Grup petit/Laboratori: 5h

Aprenentatge autònom: 3h 30m

8. Cerca semàntica i models neuronals de recuperació de la informació

Descripció:

Cerca semàntica: com fer que les cerques incloguin significats i contextos, no només paraules clau. Word embeddings i sentence embeddings. Models neuronals de recuperació de la informació. Re-ranking.

Activitats vinculades:

Activitat 2: Control 2

Activitat 3: Laboratori

Dedicació: 15h

Grup gran/Teoria: 2h

Grup petit/Laboratori: 6h

Aprenentatge autònom: 7h

SISTEMA DE QUALIFICACIÓ

L'assignatura conté els següents actes d'avaluació:

- Informes de les activitats i/o pràctiques de les sessions de laboratori (L).
- Control parcial 1 realitzat en el període d'avaluació parcial (C1).
- Control parcial 2 realitzat en el període d'avaluació final (C2).

La ponderació de les qualificacions obtingudes és la següent:

$$0.4*L + 0.3*C1 + 0.3*C2$$

NORMES PER A LA REALITZACIÓ DE LES PROVES.

Els informes de les activitats de laboratori es lliuren de forma no presencial dins el termini indicat a cada sessió.

Els controls parcials són presencials i individuals.

BIBLIOGRAFIA

Bàsica:

- 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 línia]. 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.
- Russell, Matthew A; Klassen, Mikhail. Mining the social web : data mining Facebook, Twitter, LinkedIn, Instagram, Github, and more [en línia]. 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?pg-origsite=primo&docID=561114>. ISBN 9781491973509.

RECURSOS

Altres recursos:

Enllaços 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>)