

Guia docent

270028 - CAIM - Cerca i Anàlisi d'Informació Massiva

Última modificació: 11/07/2025

Unitat responsable: Facultat d'Informàtica de Barcelona
Unitat que imparteix: 723 - CS - Departament de Ciències de la Computació.
Titulació: GRAU EN ENGINYERIA INFORMÀTICA (Pla 2010). (Assignatura optativa).
Curs: 2025 **Crèdits ECTS:** 6.0 **Idiomes:** Català, Castellà

PROFESSORAT

Professorat responsable: DAVID GARCIA SORIANO - MARTA ARIAS VICENTE

Altres: Primer quadrimestre:
MARTA ARIAS VICENTE - 11, 12, 13
DAVID GARCIA SORIANO - 11, 12

CAPACITATS PRÈVIES

Genèricament, les que s'adquireixen a les assignatures del grau que són requisits d'aquesta.

Específicament:

- Usar amb comoditat els conceptes bàsics d'àlgebra lineal, matemàtica discreta, probabilitat i estadística.
- Programar amb comoditat en llenguatges orientats a objectes, incloent 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 algorisme 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

- Pre-requisit BD
- Pre-requisit PE
- Co-requisit PROP

COMPETÈNCIES DE LA TITULACIÓ A LES QUALS CONTRIBUEIX L'ASSIGNATURA

Específiques:

CCO2.5. Implementar software de cerca d'informació (information retrieval).
CSI2.3. Demostrar coneixement i capacitat d'aplicació dels sistemes d'extracció i de gestió del coneixement.
CSI2.6. Demostrar coneixement i capacitat d'aplicació dels sistemes d'ajuda a la presa de decisions i de bussines intelligence.

Genèriques:

G7. Detectar carències en el coneixement propi i superar-les mitjançant la reflexió crítica i l'elecció de la millor actuació per ampliar aquest coneixement. Capacitat per a l'aprenentatge de nous mètodes i tecnologies, i versatilitat per a adaptar-se a noves situacions.

METODOLOGIES DOCENTS

- Classes de teoria. Abans de cada classe, l'estudiant haurà d'haver llegit els apunts o materials del tema a desenvolupar, anunciat amb temps. L'estudiant tindrà també a la seva disposició un qüestionari de preguntes bàsiques del tema, per comprovar si el grau de comprensió que ha assolit. A la classe, el professor exposarà els punts principals, donant per suposat que l'estudiant ha fet la feina indicada i intentat respondre el qüestionari, i es discutiran en comú els dubtes que puguin haver trobat els estudiants.

- Classes de problemes. Professors i estudiants comentaran i compararan les solucions dels problemes que el professor haurà indicat amb temps suficient abans de cada classe. Les discussions poden fer-se en comú entre tota la classe o particularment entre el professor i un alumne. El professor donarà per suposat que els estudiants han passat un temps raonable intentant resoldre aquests exercicis, i prioritzarà l'atenció a aquells que ho hagin fet així.

- Classes de laboratori: Abans de cada classe, l'estudiant haurà d'haver llegit el guió de la feina pràctica a desenvolupar en la sessió. Durant la classe, l'estudiant durà a terme la feina indicada en el guió amb el guiatge del professor. En moltes de les sessions, el guió contindrà feina que, probablement, calgui acabar com a treball personal després de la sessió de laboratori. Per a la majoria de sessions de laboratori caldrà redactar un informe curt de la feina feta o lliurar aquesta feina (p.ex., fitxers de resultats i programes escrits).

- Treball personal: Així doncs, cada tipus d'activitat presencial implica una certa quantitat de treball personal abans o després. Addicionalment, algun tema o temes de l'assignatura poden no tenir classes de teoria o d'exercicis associats, i els estudiants hauran d'estudiar-lo pel seu compte, i usar les sessions d'activitats dirigides si ho desitgen per avaluar que han fet el progrés suficient.

Donat que l'assignatura apareix a dues especialitats diferents, podran proposar-se activitats (a teoria, problemes i laboratori) lleugerament diferents per als estudiants d'ambdues especialitats, vetllant perquè no hi hagi greuges comparatius pel que fa a dificultat o càrrega de treball.

OBJECTIUS D'APRENTATGE DE L'ASSIGNATURA

1. Conèixer els problemes associats a l'emmagatzemament i recuperació de la informació, sobretot de tipus textual.
2. Entendre que l'efectivitat en la cerca i recuperació de la informació està molt relacionada amb l'organització i descripció d'aquesta informació.
3. Conèixer i entendre l'estructura, arquitectura i funcionament de la web, i els elements relacionats amb ella: índexos, cercadors, crawlers, entre altres.
4. Conèixer i entendre els paràmetres de descripció de xarxes complexes, així com els algorismes principals d'anàlisi de la seva estructura.
5. Reconèixer les oportunitats d'ús de la informació massiva per als fins d'una organització, i triar els mètodes, eines i procediments més adients.
6. Poder decidir les tècniques de recuperació de la informació que poden ser efectives en un sistema d'informació concret, sobretot de tipus textual.
7. Poder avaluar l'efectivitat i utilitat, d'acord amb diversos criteris, d'un sistema de recuperació de la informació.
8. Poder implementar les principals tècniques vistes a l'assignatura
9. Saber utilitzar, adaptar i estendre software obert.

HORES TOTALES DE DEDICACIÓ DE L'ESTUDIANTAT

Tipus	Hores	Percentatge
Hores grup mitjà	15,0	10.00
Hores grup petit	30,0	20.00
Hores aprenentatge autònom	84,0	56.00
Hores grup gran	15,0	10.00
Hores activitats dirigides	6,0	4.00

Dedicació total: 150 h

CONTINGUTS

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.

Models de recuperació de la informació

Descripció:

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

Implementació: Indexació i cerques

Descripció:

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

Avaluació en recuperació de la informació

Descripció:

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

Cerca a internet

Descripció:

Ranking i relevància per a models web. Algorisme PageRank. Crawling. Arquitectura de un sistema simple de cerca a la web.

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.

Anàlisi de xarxes

Descripció:

Paràmetres descriptius i característiques de les xarxes: grau, diàmetre, xarxes "small-world", entre altres. Algorismes sobre xarxes: clustering, detecció de comunitats i de nodes influents, reputació, entre altres.

Sistemes d'informació basats en l'explotació d'informació massiva. Combinació amb altres tecnologies.

Descripció:

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

ACTIVITATS

Introducció i Models de Recuperació de la Informació

Descripció:

2 hores de teoria, 2 de problemes i 4 de laboratori sobre els continguts "Introducció" i "Models de recuperació de la informació".
Vegeu descripcions a la Metodologia Docent.

Objectius específics:

1, 2, 6

Competències relacionades:

G7. Detectar carències en el coneixement propi i superar-les mitjançant la reflexió crítica i l'elecció de la millor actuació per ampliar aquest coneixement. Capacitat per a l'aprenentatge de nous mètodes i tecnologies, i versatilitat per a adaptar-se a noves situacions.

Dedicació: 25h 30m

Aprenentatge autònom: 13h 30m

Grup gran/Teoria: 4h

Grup mitjà/Pràctiques: 2h

Grup petit/Laboratori: 6h

Implementació i avaluació

Descripció:

2 hores de teoria, 2 de problemes i 4 de laboratori sobre els continguts "Implementació" i "Avaluació". Vegeu descripcions a la Metodologia Docent.

Objectius específics:

2, 7, 8, 9

Competències relacionades:

G7. Detectar carències en el coneixement propi i superar-les mitjançant la reflexió crítica i l'elecció de la millor actuació per ampliar aquest coneixement. Capacitat per a l'aprenentatge de nous mètodes i tecnologies, i versatilitat per a adaptar-se a noves situacions.

Dedicació: 14h

Aprenentatge autònom: 6h 30m

Grup gran/Teoria: 2h 30m

Grup mitjà/Pràctiques: 1h

Grup petit/Laboratori: 4h

Cerca a la web

Descripció:

2 hores de teoria, 2 de problemes i 4 de laboratori sobre el contingut "Cerca a la web".
Vegeu descripcions a la Metodologia docent.

Objectius específics:

3, 5, 9

Competències relacionades:

G7. Detectar carències en el coneixement propi i superar-les mitjançant la reflexió crítica i l'elecció de la millor actuació per ampliar aquest coneixement. Capacitat per a l'aprenentatge de nous mètodes i tecnologies, i versatilitat per a adaptar-se a noves situacions.

Dedicació: 22h

Aprenentatge autònom: 12h 30m

Grup gran/Teoria: 2h 30m

Grup mitjà/Pràctiques: 1h

Grup petit/Laboratori: 6h

Primer examen parcial

Descripció:

Examen parcial de la primera part de l'assignatura.

Objectius específics:

1, 2, 3, 5, 6, 7

Competències relacionades:

G7. Detectar carències en el coneixement propi i superar-les mitjançant la reflexió crítica i l'elecció de la millor actuació per ampliar aquest coneixement. Capacitat per a l'aprenentatge de nous mètodes i tecnologies, i versatilitat per a adaptar-se a noves situacions.

Dedicació: 10h

Aprenentatge autònom: 8h

Activitats dirigides: 2h

Arquitectura de sistemes de cerca a la web

Descripció:

2 hores de teoria i 6 de laboratori sobre els continguts "Arquitectura". Vegeu descripcions a la Metodologia Docent.

Objectius específics:

3, 6, 8, 9

Competències relacionades:

G7. Detectar carències en el coneixement propi i superar-les mitjançant la reflexió crítica i l'elecció de la millor actuació per ampliar aquest coneixement. Capacitat per a l'aprenentatge de nous mètodes i tecnologies, i versatilitat per a adaptar-se a noves situacions.

Dedicació: 23h 30m

Aprenentatge autònom: 12h 30m

Grup gran/Teoria: 3h 30m

Grup mitjà/Pràctiques: 1h 30m

Grup petit/Laboratori: 6h

Anàlisi de xarxes

Descripció:

4 hores de teoria i 6 de laboratori sobre els contingut "Anàlisi de xarxes". Vegeu descripcions a la Metodologia Docent.

Objectius específics:

4, 6, 7, 8, 9

Competències relacionades:

G7. Detectar carències en el coneixement propi i superar-les mitjançant la reflexió crítica i l'elecció de la millor actuació per ampliar aquest coneixement. Capacitat per a l'aprenentatge de nous mètodes i tecnologies, i versatilitat per a adaptar-se a noves situacions.

Dedicació: 21h 30m

Aprenentatge autònom: 12h 30m

Grup gran/Teoria: 3h 30m

Grup mitjà/Pràctiques: 1h 30m

Grup petit/Laboratori: 4h

Sistemes d'informació basats en anàlisi d'informació massiva

Descripció:

Classes de teoria, problemes i laboratori sobre aquest contingut. L'èmfasi és en la discussió de casos pràctics a classe de problemes. Vegeu descripcions a la Metodologia docent.

Objectius específics:

5, 6, 7, 9

Competències relacionades:

G7. Detectar carències en el coneixement propi i superar-les mitjançant la reflexió crítica i l'elecció de la millor actuació per ampliar aquest coneixement. Capacitat per a l'aprenentatge de nous mètodes i tecnologies, i versatilitat per a adaptar-se a noves situacions.

Dedicació: 18h 30m

Aprenentatge autònom: 12h 30m

Grup gran/Teoria: 1h 30m

Grup mitjà/Pràctiques: 0h 30m

Grup petit/Laboratori: 4h

Segon examen parcial o examen final

Descripció:

L'estudiant tria entre fer un examen de la segona meitat del curs, o bé un examen de tot el curs.

Objectius específics:

1, 2, 3, 4, 5, 6, 7, 8, 9

Competències relacionades:

G7. Detectar carències en el coneixement propi i superar-les mitjançant la reflexió crítica i l'elecció de la millor actuació per ampliar aquest coneixement. Capacitat per a l'aprenentatge de nous mètodes i tecnologies, i versatilitat per a adaptar-se a noves situacions.

Dedicació: 15h

Aprenentatge autònom: 12h

Activitats dirigides: 3h

SISTEMA DE QUALIFICACIÓ

L'assignatura comprendrà els següents actes avaluatoris:

- Informes de les activitats de laboratori, que caldrà haver lliurat dins un termini indicat per a cada sessió (orientativament, 2 setmanes). A partir d'una mitjana ponderada de les notes d'aquests informes es calcularà una nota de laboratori, L.
- Un primer examen parcial, fet cap a meitat del curs, de la matèria vista fins llavors. Sigui P1 la nota obtinguda en aquest examen.
- En el dia designat dins del període d'exàmens, l'estudiant triarà entre dues opcions exclusives: 1) fer un segon examen parcial de la matèria no coberta pel primer parcial (que, si és el cas, dóna una nota P2) o 2) fer un examen final de tota la matèria (que, si és el cas, dóna una nota F). No hi ha cap requisit de nota mínima a P1 a l'hora de fer aquesta tria.

Les quatre notes L, P1, P2 i F són entre 0 i 10. La nota final de l'assignatura serà:

$$0.4*L + \max(0.3*P1+0.3*P2, 0.6*F).$$

Triar fer l'examen final implica $P2 = 0$. Triar el segon parcial implica $F = 0$.

No presentar-se el dia de l'examen final / segon parcial implica que l'estudiant ha triat segon parcial.

Pel que fa a la nota de la competència associada a Aprenentatge Autònom, es calcularà una nota numèrica així:

- Per a l'i-èssim informe de laboratori a lliurar, el valor R_i valdrà 1 si l'informe ha estat lliurat dins del termini establert i (a criteri del professor) denota un esforç raonable de resolució de la feina corresponent, i 0 en cas contrari. Sigui Rsum la suma de tots els R_i (que pot arribar a ser k si es demanen k informes).

- Algunes de les preguntes dels exàmens finals o parcials, marcades especialment, versaran total o parcialment sobre temes que l'estudiant haurà de preparar pel seu compte, amb poca o cap cobertura a classe de teoria i problemes, que s'hauran indicat durant el curs. Sigui E la mitjana ponderada d'aquestes preguntes en els exàmens aplicables a l'estudiant, i escalada a l'interval [0,1].

Sigui S el valor de $(Rsum/k+E)/2$, que serà entre 0 i 1.

La nota de la competència serà:

- D si S és inferior a 0.5
- C si S és entre 0.5 i 0.599
- B si S és entre 0.6 i 0.799
- A si S és 0.8 o més.

BIBLIOGRAFIA

Bàsica:

- Baeza-Yates, R.; Ribeiro-Neto, B. Modern information retrieval: the concepts and technology behind search. 2nd ed. Harlow [etc.]: Addison-Wesley / Pearson, 2011. ISBN 9780321416919.
- Manning, C.D.; Raghavan, P; Schütze, H. Introduction to information retrieval. New York: Cambridge University Press, 2008. ISBN 9780521865715.
- McCandless, M.; Hatcher, E.; Gospodnetic, O. Lucene in action. 2nd ed. Greenwich, Conn: Manning, 2010. ISBN 9781933988177.
- 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 from Facebook, Twitter, and LinkedIn, Instagram, GitHub, and more. 3rd ed. Sebastopol, [California]: O'Reilly Media, 2018. ISBN 9781491973509.

RECURSOS

Enllaç web:

- <https://research.facebook.com/blog/three-and-a-half-degrees-of-separation/?refid>