Guia docent
270963 - IRRS - Recuperació de la Informació i Sistemes Recomanadors

Última modificació: 23/11/2023

Unitat responsable: Facultat d'Informàtica de Barcelona
Unitat que imparteix: 723 - CS - Departament de Ciències de la Computació.

Titulació: MÀSTER UNIVERSITARI EN CIÈNCIA DE DADES (Pla 2021). (Assignatura optativa).

Curs: 2023 **Crèdits ECTS:** 6.0 **Idiomes:** Anglès

PROFESSORAT

Professorat responsable: RAMON FERRER CANCHO

Altres: Primer quadrimestre:
RAMON FERRER CANCHO - 10

CAPACITATS PRÈVIES

Les suposades a l'ingrés al MIRI més les proporcionades per la fase de formació comú.

COMPETÈNCIES DE LA TITULACIÓ A LES QUALS CONTRIBUEIX L'ASSIGNATURA

Específiques:
CE1. Desenvolupar algoritmes eficients fonamentats en el coneixement i comprensió de la teoria de la complexitat computacional i les principals estructures de dades, dins de l'àmbit de ciència de dades
CE11. Analitzar i extreure coneixement d'informació no estructurada mitjançant tècniques de processament de llenguatge natural, mineria de textos i imatges

Genèriques:
CG2. Identificar i aplicar mètodes d'anàlisi, extracció de coneixement i visualització de dades recollides en formats molt diferents

Transversals:
CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

Bàsiques:
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.
CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.

METODOLOGIES DOCENTS

Sessions de teoria + problemes de 3 hores setmanals. Les 2 primeres hores de cada sessió són de tipus teòric, i la tercera es dedica a problemes. Per a cada sessió, l'estudiant haurà de lliurar les solucions d'alguns problemes proposats però no resolts en l'anterior.

Sessions de laboratori d'1 hora setmanal. Per a moltes de les sessions, l'estudiant haurà de lliurar un informe de la feina feta i resultats obtinguts al cap d'unes 2 setmanes.

El funcionament de cada tipus de sessió es descriu a l'apartat "Activitats".

A més, cap al final del curs els estudiants hauran de presentar davant els professors i els altres matriculats que vulguin assistir un article científic relacionat amb la temàtica de l'assignatura, en una forma similar a com es presentaria en un congrés científic. Cap a la setmana 8 de curs es farà pública una llista d'articles d'entre els quals l'estudiant podrà triar-ne un. Alternativament també pot proposar un article escollit per ell per al vist-i-plau dels professors. El dia de les presentacions s'anunciarà amb al menys 2 mesos de temps, i l'ordre i hora exactes de les presentacions amb al menys 1 setmana de temps.

OBJECTIUS D'APRENENTATGE DE L'ASSIGNATURA

1.Tècniques de cerca i tractament de la informació en entorns heterogenis
2.Recommender systems
3.Algorismes avançats per a mineria de dades

HORES TOTALS DE DEDICACIÓ DE L'ESTUDIANTAT

Tipus	Hores	Percentatge
Hores grup gran	27,0	18.00
Hores grup petit	27,0	18.00
Hores aprenentatge autònom	96,0	64.00

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. Compressió 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 pera 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 Mineria de Dades i Aprenentatge. Sistemes de recomanació.

ACTIVITATS

Desenvolupament teòric dels temes 1 a 8 de l'assignatura

Descripció:
L'alumne atendrà a l'exposició del professor i participarà activament en la discussió inicial del repte presentat.

Objectius específics:
1, 2, 3

Competències relacionades:
CG2. Identificar i aplicar mètodes d'anàlisi, extracció de coneixement i visualització de dades recollides en formats molt diferents
CE11. Analitzar i extreure coneixement d'informació no estructurada mitjançant tècniques de processament de llenguatge natural, mineria de textos i imatges
CE1. Desenvolupar algoritmes eficients fonamentats en el coneixement i comprensió de la teoria de la complexitat computacional i les principals estructures de dades, dins de l'àmbit de ciència de dades
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.
CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.

Dedicació: 52h
Grup gran/Teoria: 26h
Aprenentatge autònom: 26h

Exercicis sobre els temes 1 a 8 de l'assignatura

Descripció:
A cada sessió, el professor planteja una col·lecció de problemes (orientativament, entre 4 i 7) del tema que s'acaba de tractar teòricament. A continuació es resolen conjuntament alguns dels problemes proposats (orientativament, 3). Els estudiants han de resoldre la resta dels problemes fora d'hores de classe, i lliurar-los a l'inici de la la sessió següent. Part de la sessió es dedica a comentar conjuntament els dubtes que puguin haver sorgit en la resolució d'aquests problemes pendents de la sessió anterior.
parte de la sesión se reserva para la discusión de dudas que puedan haber surgido en la resolución de los problemas pendientes de la sesión anterior.

Objectius específics:
1, 2, 3

Competències relacionades:
CG2. Identificar i aplicar mètodes d'anàlisi, extracció de coneixement i visualització de dades recollides en formats molt diferents
CE11. Analitzar i extreure coneixement d'informació no estructurada mitjançant tècniques de processament de llenguatge natural, mineria de textos i imatges
CE1. Desenvolupar algoritmes eficients fonamentats en el coneixement i comprensió de la teoria de la complexitat computacional i les principals estructures de dades, dins de l'àmbit de ciència de dades
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.
CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.

Dedicació: 39h
Grup mitjà/Pràctiques: 13h
Aprenentatge autònom: 26h

Treball de laboratori sobre els temes 1 a 8

Descripció:
El professor planteja un treball de tipus pràctic relacionat amb els temes vistos més recentment. Aquest pot consistir en l'anàlisi d'unes dades donades (o que calgui buscar), implementar un dels algorismes vistos a classe, o proposar una solució a un cas concret de necessitat de tècniques de recuperació de la informació. L'estudiant completa tant com sigui possible el treball en l'hora de classe, encara que algun temps addicional pot ser ocasionalment necessari. En moltes de les sessions es demanarà un informe de la feina feta i els resultats obtinguts, a lliurar en el termini que es definirà en cada cas (orientativament, 2 setmanes).

Objectius específics:
1, 2, 3

Competències relacionades:
CG2. Identificar i aplicar mètodes d'anàlisi, extracció de coneixement i visualització de dades recollides en formats molt diferents
CE11. Analitzar i extreure coneixement d'informació no estructurada mitjançant tècniques de processament de llenguatge natural, mineria de textos i imatges
CE1. Desenvolupar algoritmes eficients fonamentats en el coneixement i comprensió de la teoria de la complexitat computacional i les principals estructures de dades, dins de l'àmbit de ciència de dades
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.
CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.

Dedicació: 26h
Grup petit/Laboratori: 13h
Aprenentatge autònom: 13h

Examen final

Descripció:
Examen final sobre el contingut de tota l'assignatura

Objectius específics:
1, 2, 3

Competències relacionades:
CG2. Identificar i aplicar mètodes d'anàlisi, extracció de coneixement i visualització de dades recollides en formats molt diferents
CE11. Analitzar i extreure coneixement d'informació no estructurada mitjançant tècniques de processament de llenguatge natural, mineria de textos i imatges
CE1. Desenvolupar algoritmes eficients fonamentats en el coneixement i comprensió de la teoria de la complexitat computacional i les principals estructures de dades, dins de l'àmbit de ciència de dades
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.
CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.

Dedicació: 18h
Activitats dirigides: 3h
Aprenentatge autònom: 15h

Estudi i presentació d'un article científic

Descripció:
Estudi i presentació d'un article científic relacionat amb la temàtica de l'assignatura

Objectius específics:
1, 2, 3

Competències relacionades:
CG2. Identificar i aplicar mètodes d'anàlisi, extracció de coneixement i visualització de dades recollides en formats molt diferents
CE11. Analitzar i extreure coneixement d'informació no estructurada mitjançant tècniques de processament de llenguatge natural, mineria de textos i imatges
CE1. Desenvolupar algoritmes eficients fonamentats en el coneixement i comprensió de la teoria de la complexitat computacional i les principals estructures de dades, dins de l'àmbit de ciència de dades
CT5. TERCERA LLENGUA: Conèixer una tercera llengua, preferentment l'anglès, amb un nivell adequat oral i escrit i en consonància amb les necessitats que tindran els titulats i titulades.

CT4. ÚS SOLVENT DELS RECURSOS D'INFORMACIÓ: Gestionar l'adquisició, l'estructuració, l'anàlisi i la visualització de dades i informació de l'àmbit d'especialitat, i valorar de forma crítica els resultats d'aquesta gestió.
CB6. Que els estudiants sàpiguen aplicar els coneixements adquirits y la seva capacitat de resolució de problemes en entorns nous o poc coneguts dins de contexts més amplis (o multidisciplinaris) relacionats amb la seva àrea d'estudi.
CB10. Posseir i comprendre coneixements que aportin una base o oportunitat de ser originals en el desenvolupament i/o aplicació d'idees, sovint en un context de recerca.
CB7. Que els estudiants siguin capaços d'integrar coneixements i enfrontar-se a la complexitat de formular judicis a partir d'una informació que, essent incomplerta o limitada, inclogui reflexions sobre les responsabilitats socials i ètiques vinculades a l'aplicació dels seus coneixements i judicis.

Dedicació: 13h
Activitats dirigides: 3h
Aprenentatge autònom: 10h

SISTEMA DE QUALIFICACIÓ

Siguin:

- NF la nota de l'examen final,
- NE la nota dels lliuraments d'exercicis,
- NL la nota de les pràctiques de laboratori,
- NA la nota de la presentació d'un article científic,

totes en el rang de 0 a 10.

La nota final de l'assignatura és 0.3*NF + 0.25*NL + 0.25*NE + 0.2*NA.

BIBLIOGRAFIA

Bàsica:
- Baeza-Yates, R.; Ribeiro-Neto, B. Modern information retrieval: the concepts and technology behind search. 2nd ed. Addison-Wesley / Pearson, 2011. ISBN 9780321416919.
- 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. Boston [etc.]: Pearson, 2010. ISBN 9780131364899.
- Russell, M.A.; Klassen, M. Mining the social web: data mining Facebook, Twitter, LinkedIn, Instagram, Github, and more. 3rd ed. Sebastopol, [California]: O'Reilly Media, 2018. ISBN 9781491973509.
- McCandless, M.; Hatcher, E.; Gospodnetic, O. Lucene in action. 2nd ed. Greenwich, Conn: Manning, 2010. ISBN 9781933988177.

RECURSOS

Enllaç web:
- http://www.cs.upc.edu/~IR-MIRI/. Supporting web of the course