

Guia docent

340459 - PEDT - Processament i Explotació de Dades Textuals

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 Roig

Altres:

CAPACITATS PRÈVIES

- Saber aplicar els conceptes bàsics d'àlgebra lineal, matemàtica discreta, probabilitat i estadística.
- Tenir coneixements bàsics de programació amb Python.
- Tenir coneixements bàsics dels conceptes estudiats en les assignatures de MIDA i REIN.

REQUISITS

Haver aprovat ESTA, AMEP i DABD o almenys estar-hi matriculat/da.
Es recomana haver cursat MIDA i REIN.

COMPETÈNCIES DE LA TITULACIÓ A LES QUALS CONTRIBUEIX L'ASSIGNATURA

Específiques:

I_CECO7. 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.

I_CECO1. CECO1. Capacitat per a tenir un coneixement profund dels principis fonamentals i models de la computació i saber-los aplicar per a interpretar, seleccionar, valorar, modelar, i crear nous conceptes, teories, usos i desenvolupaments tecnològics relacionats amb la informàtica.

I_CECO4. CECO4. Capacitat per a conèixer els fonaments, paradigmes i tècniques pròpies dels sistemes intel·ligents i analitzar, dissenyar i construir sistemes, serveis i aplicacions informàtiques que utilitzin aquestes tècniques en qualsevol àmbit d'aplicació.

I_CEIS6. 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.

I_CEIS4. 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.

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 del Processament i Explotació de Dades Textuals (Text Mining).

OBJECTIUS D'APRENENTATGE DE L'ASSIGNATURA

Les dades textuais es troben arreu per exemple en llibres, articles, pàgines web, lleis, anàlisis financeres, registres mèdics, xarxes socials, etc. S'estima que representen entre el 80% i el 90% de les dades emmagatzemades. Per poder extraure, resumir i analitzar informació a partir de grans volums de dades textuais es requereixen mètodes específics. El camp conegut com a Text Mining fa ús de tècniques de computació per extraure informació de dades textuais de forma automàtica.

L'assignatura cobreix els components bàsics del Processament del Llenguatge Natural (Natural Language Processing) i com s'usen en les tasques de Mineria de Textos (Text Mining). També s'estudien algunes de les seves aplicacions com ara Classificació de Documents, Anàlisi de Sentiment (Sentiment Analysis o Opinion Mining) i Extracció d'Informació.

HORES TOTALES DE DEDICACIÓ DE L'ESTUDIANTAT

Tipus	Hores	Percentatge
Hores grup petit	30,0	50.00
Hores grup gran	30,0	50.00

Dedicació total: 60 h

CONTINGUTS

1. Processos d'obtenció de dades textuais

Descripció:

Obtenir o construir un corpus. Creació d'un corpus amb dades extretes de fonts diverses: correus electrònics, articles de la Wikipedia, informes financers, obres literàries o pàgines web d'interès.

Rastreig de les dades (scrapping o web crawling).

Activitats vinculades:

LABORATORI

QÜESTIONARIS

PROJECTE

Dedicació: 6h 30m

Grup gran/Teoria: 4h

Grup petit/Laboratori: 2h 30m

2. Preprocessament de dades textuais

Descripció:

Processament sintàctic simple: neteja del text, normalització i tokenització.

Processament lingüístic avançat: desambiguació i etiquetat gramatical (part-of-speech tagging).

Activitats vinculades:

LABORATORI

QÜESTIONARIS

PROJECTE

Dedicació: 7h 30m

Grup gran/Teoria: 4h

Grup petit/Laboratori: 3h 30m

3. Processament del Llenguatge Natural: tasques principals i aplicacions

Descripció:

Introducció al PLN. Tasques principals: etiquetat gramatical (PoS tagging), anàlisi sintàctica i interpretació semàntica. Algunes aplicacions (demos): classificació de documents, clustering de documents, anàlisi de sentiment, extracció d'informació, resum automàtic, traducció automàtica.

Activitats vinculades:

LABORATORI
QÜESTIONARIS
PROJECTE

Dedicació: 13h

Grup gran/Teoria: 6h

Grup petit/Laboratori: 7h

4. Eines de PLN i repositoris de dades per la mineria de textos

Descripció:

Eines de PLN amb Python: Scikit-Learn, Natural Language Toolkit (NLTK), Gensim, spaCy, NetworkX.
Repositoris de dades per la mineria de textos accessibles on-line.

Activitats vinculades:

LABORATORI
QÜESTIONARIS
PROJECTE

Dedicació: 12h

Grup gran/Teoria: 6h

Grup petit/Laboratori: 6h

5. Introducció a les xarxes neuronals i l'aprenentatge profund (Deep Learning) aplicat a la mineria de textos

Descripció:

Vectorització dels textos: bag-of-words, tf-idf, word embeddings.
Xarxes neuronals. Aplicacions de l'aprenentatge profund (Deep Learning) a la mineria de textos.

Activitats vinculades:

LABORATORI
QÜESTIONARIS
PROJECTE

Dedicació: 12h

Grup gran/Teoria: 6h

Grup petit/Laboratori: 6h

6. Presentacions dels projectes

Descripció:

Presentacions dels projectes elaborats pels estudiants.

Activitats vinculades:

PROJECTE

Dedicació: 9h

Grup gran/Teoria: 4h

Grup petit/Laboratori: 5h

SISTEMA DE QUALIFICACIÓ

- Avaluació de les activitats realitzades en les sessions de laboratori: 60%
- Realització i presentació pública d'un treball d'anàlisi sobre un dels temes estudiats en el curs: 30%
- Qüestionaris: 10%

Atès que el 100% de l'assignatura s'avalua amb treballs pràctics, no hi ha cap control final de caràcter global ni tampoc cap control de reavaluació en forma d'examen escrit.

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 qüestionaris són presencials i individuals.

BIBLIOGRAFIA

Bàsica:

- Jurafsky, Dan; Martin, James H. Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition [en línia]. 3rd ed. Upper Saddle River: Els autors, 2019 [Consulta: 28/04/2022]. Disponible a: <https://web.stanford.edu/~jurafsky/slp3/>.
- Ignatow, Gabe; Mihalcea, Rada F. An Introduction to text mining : research design, data collection, and analysis. Thousand Oaks, California: SAGE Publications, Inc, 2018. ISBN 9781506337005.