

**FITXA IDENTIFICATIVA****Dades de l'Assignatura**

Codi	36429
Nom	Processament del llenguatge natural
Cicle	Grau
Crèdits ECTS	6.0
Curs acadèmic	2022 - 2023

Titulació/titulacions

Titulació	Centre	Curs	Període
1400 - Grau d'Enginyer Informàtica	Escola Tècnica Superior d'Enginyeria	4	Segon quadrimestre
1406 - Grau en Ciència de Dades	Escola Tècnica Superior d'Enginyeria	3	Segon quadrimestre

Matèries

Titulació	Matèria	Caràcter
1400 - Grau d'Enginyer Informàtica	16 - Matèria Optativa	Optativa
1406 - Grau en Ciència de Dades	9 - Aprenentatge automàtic i mineria de dades	Obligatòria

Coordinació

Nom	Departament
VILA FRANCES, JOAN	242 - Enginyeria Electrònica

RESUM

Actualment, gran part de les dades disponibles per a l'anàlisi estan formats per informació no estructurada en forma de textos en llenguatge natural. Entre aquesta informació trobem pàgines web (Wikipedia, periòdics digitals, blogs) o xarxes socials (Facebook, Twitter). Poder analitzar aquests textos, mitjançant algorismes de processament de llenguatge natural (PLN), resulta molt útil perquè les organitzacions puguin prendre millors decisions.

Els algorismes d'aprenentatge automàtic no són capaços d'entendre text o caràcters, per la qual cosa el PLN realitza tot el processament necessari per a convertir aquestes dades en forma de text en un format comprensible per les màquines (números) i així poder realitzar tot tipus d'anàlisi posterior. Entre les aplicacions més comunes del PLN es troben la classificació de textos, cerca i extracció d'informació, traducció automàtica o sistemes de resposta automàtica, entre altres.



Tots els passos del PLN, des de la captura del text en qualsevol format a la manipulació i anàlisi d'aquest per a obtenir la informació rellevant, són abordats en l'assignatura obligatòria 36429, Processat de Llenguatge Natural que s'imparteix en el segon quadrimestre del tercer curs.

Les classes de teoria s'impartiran en castellà i les classes pràctiques i de laboratori segons consta en la fitxa de l'assignatura disponible en la web del grau.

CONEIXEMENTS PREVIS

Relació amb altres assignatures de la mateixa titulació

No heu especificat les restriccions de matrícula amb altres assignatures del pla d'estudis.

Altres tipus de requisits

Es recomana haver superat les assignatures de Models Lineals (segon curs) i Aprenentatge Màquina (primer quadrimestre del tercer curs)

COMPETÈNCIES

1400 - Grau d'Enginyer Informàtica

- C1 - Capacitat per conèixer els fonaments, els paradigmes i les tècniques propis dels sistemes intel·ligents, i analitzar, dissenyar i construir sistemes, serveis i aplicacions informàtiques que utilitzen aquestes tècniques en qualsevol àmbit d'aplicació.
- C2 - Capacitat per adquirir, obtenir, formalitzar i representar el coneixement humà en una forma computable per a la resolució de problemes mitjançant un sistema informàtic en qualsevol àmbit d'aplicació, particularment els relacionats amb aspectes de computació, percepció i actuació en ambients o entorns intel·ligents.
- C3 - Capacitat per conèixer i desenvolupar tècniques d'aprenentatge computacional i dissenyar i implementar aplicacions i sistemes que les utilitzen, incloent-hi les dedicades a extracció automàtica d'informació i de coneixement a partir de grans volums de dades.

1406 - Grau en Ciència de Dades

- (CG06) Capacitat d'accés i gestió de la informació en diferents formats per a la seva posterior anàlisi amb la finalitat d'obtenir coneixement a partir de dades.
- (CT04) Ser responsables del seu propi desenvolupament professional i de la seva especialització, aplicant els coneixements adquirits en la identificació de sortides professionals i jaciments d'ocupació.
- (CE03) Capacitat per resoldre problemes de classificació, modelització, segmentació i predicció a partir d'un conjunt de dades.



- (CE07) Capacitat per modelar la dependència entre una variable resposta i diverses variables explicatives, en conjunts de dades complexes, mitjançant tècniques d'aprenentatge màquina, interpretant els resultats obtinguts.
- (CB5) Que els estudiants hagen desenvolupat aquelles habilitats d'aprenentatge necessàries per a emprendre estudis posteriors amb un alt grau d'autonomia.

RESULTATS DE L'APRENENTATGE

Saber segmentar text en elements simples (CG06)

Conèixer les tècniques de processament de llenguatge natural (CG06, CT04, CE03, CE07).

Conèixer e implementar les aplicacions més esteses de processament de llenguatge natural (CB05, CT05, CE03, CE07).

Com a conseqüència dels resultats de l'aprenentatge adquirits, els estudiants adquiriran les següents habilitats:

- Ser capaç de carregar i tractar text en Python.
- Saber usar expressions regulars sobre text.
- Conèixer i saber usar les llibreries més importants de PLN per a Python.
- Ser capaç de convertir text en vectors numèrics per al seu posterior tractament amb algorismes de aprenentatge màquina.
- Ser capaç de dissenyar una aplicació de classificació de text.
- Ser capaç d'analitzar grans volums de text per a extrau les temàtiques més representatives i realitzar recerca d'informació.

DESCRIPCIÓ DE CONTINGUTS

1. Introducció al Processament de Llenguatge Natural

- 1.1. Què és el PLN
- 1.2. La importància del text
- 1.3. Aproximacions històriques al PLN
- 1.4. Aplicacions i flux de treball



2. Ús de text en Python

- 2.1. Cadenes de text en Python
- 2.2. Expressions regulars
- 2.3. Càrrega de text
- 2.4. Captura de contingut web (web scraping)
- 2.5. Llibreries PLN en Python

3. Pre-processament de text

- 3.1. Divisió de text
- 3.2. Neteja i normalització del text
- 3.3. Anàlisi morfològic
- 3.4. Anàlisi semàntic
- 3.5. Anàlisi gramatical

4. Extracció de característiques

- 4.1. Característiques simples
- 4.2. Model Bag of Words
- 4.3. Model TF-IDF
- 4.4. Vectors de paraula (word embeddings)
- 4.5. Vectors de document

5. Aplicacions del PLN

- 5.1. Classificació
- 5.2. Extracció de la informació
- 5.3. Minería de text
- 5.4. Recerca de informació
- 5.5. Models seqüencials

6. Pràctiques de Processament de Llenguatge Natural

En este bloc es realitzaran una sèrie de exercicis pràctics per aplicar els conceptes de PLN en el aula informàtica.

Pràctica 0: Ús de text en Python (no presencial)

Pràctica 1: Expressions regulars

Pràctica 2: Web scraping

Pràctica 3: Ús de llibreries PLN en Python

Pràctica 4: Pre-processament de text en Python

Pràctica 5: Classificació de text

Pràctica 6: Extracció d'informació

**VOLUM DE TREBALL**

ACTIVITAT	Hores	% Presencial
Classes de teoria	30,00	100
Pràctiques en laboratori	20,00	100
Pràctiques en aula	10,00	100
Elaboració de treballs en grup	10,00	0
Elaboració de treballs individuals	10,00	0
Estudi i treball autònom	15,00	0
Lectures de material complementari	5,00	0
Preparació d'activitats d'avaluació	10,00	0
Preparació de classes de teoria	10,00	0
Preparació de classes pràctiques i de problemes	10,00	0
Resolució de casos pràctics	15,00	0
Resolució de qüestionaris on-line	5,00	0
TOTAL	150,00	

METODOLOGIA DOCENT

Les classes combinaran el contingut teòric amb el pràctic

MD1 - Activitats teòriques. Desenvolupament expositiu de la matèria amb la participació de l'alumnat en la resolució de qüestions puntuals. Realització de qüestionaris individuals d'avaluació.

En les activitats teòriques de caràcter presencial es desenvoluparan els temes de l'assignatura proporcionant una visió global i integradora, analitzant amb major detall els aspectes clau i de major complexitat, fomentant, en tot moment, la participació de l'alumnat (CB05, CT05).

MD2 - Activitats pràctiques. Aprenentatge mitjançant resolució de problemes, exercicis i casos d'estudi a través dels quals s'adquireixen competències sobre els diferents aspectes teòrics de la matèria. (CB05, CG06, CE03, CE07)

Les activitats teòriques es complementen amb pràctiques de laboratori amb l'objectiu de posar en ús els conceptes bàsics i ampliar-los amb el coneixement i l'experiència que es vagen adquirint durant la realització dels treballs proposats.

MD4 -Treballs en laboratori i/o aula ordenador. Aprenentatge mitjançant la realització d'activitats guiades desenvolupades de manera individual o en grups reduïts i dutes a terme en laboratoris i/o aules d'ordinador. (CB05, CG06, CT04, CE03, CE07)



A més de les activitats presencials, els estudiants hauran de fer tasques personals (fora de l'aula) sobre: qüestions i problemes, així com la preparació de classes i exàmens (estudi). Aquestes tasques es realitzaran principalment de manera individual, amb la finalitat de potenciar el treball autònom, però addicionalment s'inclouran treballs, especialment la preparació i resolució de pràctiques de laboratori, que requerisquen la participació de xicotets grups d'estudiants (2-3) per a fomentar la capacitat d'integració en grups de treball.

S'utilitzarà la plataforma d'e-learning (Aula Virtual) de la Universitat de València com a suport de comunicació amb l'alumnat. A través d'ella es tindrà accés al material didàctic utilitzat en classe i els guions de les pràctiques de laboratori, així com els problemes i exercicis a resoldre.

AVALUACIÓ

L'avaluació de l'aprenentatge dels coneixements i competències aconseguides pels estudiants es farà de forma continuada al llarg del curs, i constarà dels següents blocs:

- SE1 - Prova objectiva, consistent en un examen que consta tant de qüestions teoricopràctiques com de problemes (avaluació de competències CB05, CT05, CE03, CE07) (50%) (Nota: Tots els percentatges estan referits a la nota final)
 - SE1-1 (40%) Examen de teoria-problemes
 - SE1-2 (10%) Examen de laboratori
- SE2 - Avaluació de les pràctiques de laboratori a partir de l'elaboració de treballs/memòries i/o exposicions orals (avaluació de competències CB05, CG06, CT04, CE03, CE07) (30%)
 - SE2-1 (20%) Realització d'un mini projecte consistent en el desenvolupament d'una aplicació completa de PLN per a la classificació de textos
 - SE2-2 (10%) Assistència i avaluació de les sessions de laboratori (Activitat NO RECUPERABLE)
- SE3 - Avaluació contínua de cada alumne. (20%)
 - SE3-1 (5%) Assistència regular a les activitats presencials previstes (avaluació de competències CB05). (Activitat NO RECUPERABLE)
 - SE3-2 (15%) Resolució de qüestions i problemes proposats (avaluació de competències CB05, CG06, CE03). (Activitat NO RECUPERABLE)



La nota final de l'assignatura es calcularà com la mitjana ponderada de cadascun dels apartats anteriors, d'acord amb el següent criteri: ES-1 (50%), ES-2 (30%), ES-3 (20%).

Consideracions particulars sobre l'avaluació:

- És necessari obtenir una qualificació mínima de 4 (sobre 10) en cadascun dels apartats d'avaluació SE1-1, SE1-2 i SE2-1.
- Les activitats SE2-2, SE3-1 i SE3-2 no són recuperables.

En qualsevol cas, el sistema d'avaluació es regirà pel que s'estableix en el Reglament d'Avaluació i Qualificació de la Universitat de València per a Graus i Màsters (<https://webges.uv.es/uvTaeWeb/MuestraInformacionEdictoPublicoFrontAction.do?accion=inicio&idEdictoSeleccionado=5639>)

REFERÈNCIES

Bàsiques

- Sohom Ghosh, Dwight Gunning. Natural Language Processing Fundamentals. Packt Publishing, 2019.
- Akshay Kulkarni, Adarsha Shivananda. Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python. Apress, 2019 (disponible e-libro)
- Dipanjan Sarkar. Text Analytics with Python: A Practitioner's Guide to Natural Language Processing. Apress 2019 (disponible e-libro)
- Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python. O'Really Media, 2009.

Complementàries

- Jacob Eisentein. Natural Language Processing. 2018 (disponible bajo licencia CC-BY-NC-ND)
- Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana. Practical Natural Language Processing. O'Really Media, 2020