FITXA IDENTIFICATIVA

DADES DE L'ASSIGNATURA

Codi: 36429
Nom: Processament del llenguatge natural
Cicle: Grau
Crèdits ECTS: 6
Curs acadèmic: 2026-27

TITULACIONS

Titulació	Centre	Curs	Periode
1400 - Grau Eng.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 Eng.Informàtica	Matèria Optativa	OPTATIVA
1406 - Grau en Ciència de Dades	Aprenentatge automàtic i mineria de dades	OBLIGATÒRIA

COORDINACIÓ

PELLICER VALERO OSCAR JOSE

RESUM

El Processament del Llenguatge Natural (PLN) és una disciplina en la intersecció de la intel·ligència artificial, la informàtica i la lingüística, l'objectiu de la qual és dotar les màquines de la capacitat de comprendre, interpretar i generar llenguatge humà. Aquesta assignatura ofereix un recorregut complet pel camp, partint dels fonaments teòrics i pràctics fins a arribar a les tècniques més avançades que defineixen l'estat de l'art.

S'hi abordaran les fases clau d'un projecte de PLN, incloent-hi l'adquisició, la neteja i el preprocessament de text. S'estudiaran les tècniques fonamentals de representació de text, des dels models clàssics com Bag of Words i TF-IDF fins als embeddings semàntics que capturen el significat contextual.

El curs aborda també les arquitectures de Deep Learning que han revolucionat el camp, amb una anàlisi del mecanisme d'atenció i l'arquitectura Transformer. A partir d'ací, s'explorarà el funcionament i l'aplicació dels Grans Models de Llenguatge (Large Language Models), com els de la família GPT. S'estudiaran tècniques avançades com el fine-tuning eficient (QLoRA), l'enginyeria de prompts, i la construcció de

sistemes complexos com els de Generació Augmentada per Recuperació (RAG).

A través de sessions teòriques i pràctiques, l'alumnat adquirirà les competències necessàries per a dissenyar i implementar solucions a problemes reals com la classificació de textos, l'extracció d'informació, els sistemes de pregunta-resposta i els agents conversacionals.

CONEIXEMENTS PREVIS

RELACIÓ AMB ALTRES ASSIGNATURES DE LA MATEIXA TITULACIÓ

No s'ha especificat restriccions de matrícula amb altres assignatures del pla d'estudis.

ALTRES TIPUS DE REQUISITS

Per a un seguiment adequat de l'assignatura, es recomana posseir els coneixements següents:
- Nivell intermedi de programació en Python.
- Fonaments de Machine Learning: Comprensió dels conceptes bàsics d'aprenentatge supervisat i no supervisat (classificació, regressió, clustering) i de les mètriques d'avaluació de models.
- Coneixements bàsics d'àlgebra lineal i probabilitat: Familiaritat amb vectors, matrius i conceptes de probabilitat, necessaris per a comprendre el funcionament dels models.
- Experiència prèvia amb llibreries de l'ecosistema de Ciència de Dades en Python, com pandas, numpy i scikit-learn.

COMPETÈNCIES / RESULTATS D' APRENENTATGE

1400 - Grau Eng.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

(CB5) Que els estudiants hagen desenvolupat aquelles habilitats d'aprenentatge necessàries per a emprendre estudis posteriors amb un alt grau d'autonomia.

(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.

(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.

DESCRIPCIÓ DE CONTINGUTS

1. Fonaments i eines del processament del llenguatge natural

Aquesta unitat introductòria estableix les bases conceptuals i pràctiques de l'assignatura. S'explorarà la definició i la història del Processament del Llenguatge Natural (PLN), analitzant la complexitat inherent del llenguatge humà i els seus components, des de la morfologia fins a la pragmàtica. A nivell pràctic, s'abordarà el maneig de text en Python, incloent-hi l'ús avançat de cadenes de text i el poder de les expressions regulars per a la manipulació i cerca de patrons. Finalment, s'introduiran tècniques per a l'adquisició de corpus textuals, com el web scraping, i es presentarà l'ecosistema de llibreries i frameworks que s'utilitzaran al llarg del curs, establint les bases per al cicle de vida complet d'un projecte de PLN.

2. Enginyeria de text i anàlisi lingüística

L'objectiu d'aquesta unitat és aprendre a transformar text cru i no estructurat en dades netes i enriquides, llestes per a ser processades per algorismes. S'estudiaran en profunditat les tècniques de neteja i normalització, com la conversió a minúscules i el tractament de caràcters especials. S'abordarà la tasca fonamental de la segmentació (tokenization), tant a nivell de paraula com de sub-paraula, crucial per als models moderns. El contingut s'endinsa després en l'anàlisi lingüística computacional, cobrint l'anàlisi morfològica per a la reducció de paraules a la seua arrel o lema (stemming i lemmatization), l'etiquetatge gramatical (Part-of-Speech tagging) per a identificar la funció de cada paraula, l'anàlisi de dependències sintàctiques i el Reconeixement d'Entitats Anomenades (NER).

3. Representació vectorial del text

Aquesta unitat se centra en el pas crític de convertir el text en representacions numèriques que una màquina puga entendre, explorant l'evolució des de models lèxics a semàntics. Es començarà amb les representacions sparse (disperses) clàssiques, com el model Bag-of-Words (BoW) i el seu refinament mitjançant la ponderació TF-IDF, entenent les seues fortaleses i limitacions. A continuació, es produirà un salt conceptual cap a les representacions denses amb els Word Embeddings. S'analitzaran en detall els algorismes fundacionals com Word2Vec, GloVe i FastText, que van revolucionar el camp en ser capaços de capturar relacions semàntiques i contextuals en un espai vectorial continu, permetent realitzar operacions algebraiques amb el significat de les paraules.

4. Modelatge i aplicacions amb vectors de text

Una vegada que el text ha sigut convertit en vectors significatius, aquesta unitat explora com utilitzar aquestes representacions per a resoldre problemes concrets mitjançant algorismes de Machine Learning. S'abordarà la classificació de documents, una de les aplicacions més comunes del PLN, amb casos d'ús com l'anàlisi de sentiment. S'aprofundirà en la mineria de text no supervisada a través del modelatge de tòpics (Topic Modeling), utilitzant tècniques com LSI i LDA per a descobrir els temes latents en grans col·leccions de documents. Finalment, s'estudiaran els fonaments dels sistemes de recuperació d'informació (Information Retrieval), des dels algorismes de ranking clàssics com BM25 fins als enfocaments moderns de cerca semàntica basats en la similitud d'embeddings.

5. Arquitectures de Deep Learning i models contextuals

Aquesta unitat marca la transició cap a les arquitectures de xarxes neuronals profundes que defineixen l'estat de l'art actual. Es presentarà el context històric amb les Xarxes Neuronals Recurrents (RNN, LSTM) i Convolucionals (CNN) per al tractament de seqüències. El focus principal recaurà en l'arquitectura Transformer, desgranant el seu component més innovador: el mecanisme d'atenció (self-attention). S'analitzarà com aquesta arquitectura permet processar el text de forma no seqüencial, capturant relacions a llarga distància. Com a aplicació principal d'aquestes idees, s'estudiarà en profunditat el model BERT (Bidirectional Encoder Representations from Transformers).

6. Grans Models de Llenguatge (LLMs) i aplicacions generatives

Construint sobre la base de l'arquitectura Transformer, aquesta unitat se submergeix per complet en el paradigma dels Grans Models de Llenguatge (LLMs). S'analitzarà l'arquitectura decoder-only característica de la família de models GPT i es detallarà el seu cicle de vida: des del pre-entrenament a escala massiva fins a les fases d'ajust fi mitjançant instruction tuning i aprenentatge per reforç amb retroalimentació humana (RLHF). A nivell pràctic, s'exploraran les tècniques per a interactuar i adaptar aquests models, com l'enginyeria de prompts (prompt engineering), la creació de sistemes de Generació Augmentada per Recuperació (RAG) per a basar les respostes en coneixement extern, i les metodologies de fine-tuning eficient com LoRA i QLoRA. Finalment, s'aplicaran aquests coneixements al desenvolupament d'aplicacions generatives avançades com la traducció automàtica, el resum de text abstractiu i els agents conversacionals.

7. Laboratori de Processament del Llenguatge Natural

Aquest bloc pràctic està dissenyat per a aplicar de manera tangible els conceptes teòrics desenvolupats al llarg del curs. Les pràctiques planificades són:
Pràctica 0: Fonaments de Python per a PLN (no presencial)
Pràctica 1: Manipulació de text amb expressions regulars
Pràctica 2: Adquisició de dades textuals: Web Scraping
Pràctica 3: Anàlisi i preprocessament lingüístic
Pràctica 4: Modelatge i classificació de documents
Pràctica 5: Aplicacions amb grans models de llenguatge (LLMs)
Pràctica 6: Construcció de sistemes amb LLMs

VOLUM DE TREBALL (HORES)

ACTIVITATS PRESENCIALS

Activitat	Hores
Teoria	30,00
Pràctiques a l'aula	10,00
Laboratori	20,00
Total hores	**60,00**

ACTIVITATS NO PRESENCIALS

Activitat	Hores
Assistència a altres activitats	0,00
Elaboració de treballs individuals o en grup	20,00
Estudi i treball autònom	15,00
Preparació de classes	25,00
Preparació d'activitats d'avaluació	10,00
Resolució de casos pràctics	20,00
Total hores	**90,00**

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 per l'estudiantat 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) (35%)

- 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 (15%) Assistència i avaluació de les sessions de laboratori (Activitat NO RECUPERABLE)

- SE3 - Avaluació contínua de cada alumne. (15%)

- SE3-1 (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: SE-1 (50%), SE-2 (35%), SE-3 (15%).

Consideracions particulars sobre l'avaluació:

- Per a aprovar l'assignatura, és necessari obtindre una qualificació mínima de 5 (sobre 10) en cadascun dels apartats d'avaluació SE1-1, SE1-2 i SE2-1.

- Les activitats SE2-2 i SE3-1 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)

La còpia o plagi manifest de qualsevol activitat que forma part de l'avaluació suposarà la impossibilitat de superar l'assignatura, sotmetent-se seguidament als procediments disciplinaris oportuns indicats en el PROTOCOL D'ACTUACIÓ DAVANT PRÀCTIQUES FRAUDULENTES A LA UNIVERSITAT DE VALÈNCIA (ACGUV 123/2020).

BIBLIOGRAFIA

Jurafsky, D., & Martin, J. H. (2025). Speech and Language Processing (3rd ed. draft). https://web.stanford.edu/~jurafsky/slp3/

Lane, H., & Dyshel, M. (2025). Natural language processing in action. Simon and Schuster

Bird, S., Klein, E., & Loper, E. (2009). Natural language processing with Python: analyzing text with the natural language toolkit. "O'Reilly Media, Inc."

Vaswani, A., et al. (2017). "Attention Is All You Need". Advances in Neural Information Processing Systems 30 (NIPS 2017)

Devlin, J., et al. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. In Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics (pp. 4171-4186)

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., ... & Lowe, R. (2022). Training language models to follow instructions with human feedback. Advances in neural information processing systems, 35, 27730-27744

Dettmers, T., Pagnoni, A., Holtzman, A., & Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. Advances in neural information processing systems, 36, 10088-10115