

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

Codi	46578
Nom	Processament del llenguatge natural
Cicle	Màster
Crèdits ECTS	6.0
Curs acadèmic	2023 - 2024

Titulació/titulacions

Titulació	Centre	Curs	Període
2262 - M.U. en Ciència de Dades	Escola Tècnica Superior d'Enginyeria	1	Segon quadrimestre

Matèries

Titulació	Matèria	Caràcter
2262 - M.U. en Ciència de Dades	10 - Processament del llenguatge natural	Obligatòria

Coordinació

Nom	Departament
VIVES GILABERT, YOLANDA	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 pre-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 per obtenir la informació rellevant. Entre les aplicacions més comunes del PLN es troben la classificació de textos, cerca i extracció d'informació, sumarització, traducció automàtica o sistemes de resposta automàtica, entre altres.



Es tracta d'una assignatura teòrico-pràctica, on hi haurà sessions de teoria, sessions pràctiques i sessions mixtes. Les classes de teoria s'impartiran en castell

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 l'assignatura d'Aprenentatge Màquina (primer quadrimestre).

COMPETÈNCIES

2262 - M.U. en Ciència de Dades

- Que els estudiants sàpiguem comunicar les conclusions (i els coneixements i les raons últimes que les sustenten) a públics especialitzats i no especialitzats d'una manera clara i sense ambigüitats.
- Que els estudiants posseïsquen les habilitats d'aprenentatge que els permeten continuar estudiant d'una forma que haurà de ser en gran manera autodirigida o autònoma.
- 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.
- Ser capaços de valorar la necessitat de completar la seua formació tècnica, científica, en llengües, en informàtica, en literatura, en ètica, social i humana en general, i d'organitzar el seu propi autoaprenentatge amb un alt grau d'autonomia
- Capacitat per a treballar en equip per a arribar a solucions de problemes interdisciplinaris usant tècniques d'anàlisi de dades.
- Ser capaços d'accedir a ferramentes d'informació (bibliogràfiques i d'ocupació) i utilitzar-les apropiadament.
- Ser capaços d'assumir la responsabilitat del seu propi desenvolupament professional i de la seua especialització en un o més camps d'estudi, aplicant els coneixements adquirits en la identificació d'eixides professionals i jaciments d'ocupació.
- Extraure coneixement de conjunts de dades en diferents formats.
- Modelar la dependència entre una variable resposta i diverses variables explicatives, en conjunts de dades complexes, per mitjà de tècniques d'aprenentatge màquina, interpretant els resultats obtinguts.
- Usar les tècniques de processament de llenguatge natural per analitzar textos extraient-ne coneixement útil.



RESULTATS DE L'APRENTATGE

Aprendre les diferents tècniques d'extracció i processament de la informació sobre dades no estructurades de text.

Saber transformar text en vectors numèrics (embeddings) per al seu ús en algoritmes d'aprenentatge automàtic.

Conèixer i saber aplicar models de DL sobre text, especialment els models basats en atenció (transformers).

Conèixer les diferents aplicacions del NLP en classificació, extracció de coneixement i generació de text.

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. Preprocessament 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. Mineria de text
- 5.4. Recerca de informació
- 5.5. Models seqüencials

6. Aprenentatge profund en PLN

- 6.1 Xarxes neuronals recurrents en PLN
- 6.2 Atenció en aprenentatge profund
- 6.3 Transformers
- 6.4 Aplicacions

VOLUM DE TREBALL

ACTIVITAT	Hores	% Presencial
Pràctiques en laboratori	28,00	100
Classes de teoria	28,00	100
Classes teoricopràctiques	4,00	100
Assistència a esdeveniments i activitats externes	2,00	0
Elaboració de treballs en grup	15,00	0
Elaboració de treballs individuals	3,00	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

Activitats teòriques.



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

Activitats pràctiques.

- Aprenentatge mitjançant la resolució de problemes, exercicis i casos d'estudi a través dels quals s'adquireixen competències sobre els diferents aspectes de la matèria.

Treballs en laboratori i/o aula d'ordinadors.

- Aprenentatge mitjançant la realització d'activitats desenvolupades de forma individual o en grups reduïts i portades a terme en aules d'ordinadors.

AVALUACIÓ

Prova objectiva, consistent en un o diversos exàmens que constaran tant de qüestions teòrico-pràctiques com de problemes (entre el 20% i el 80%).

Avaluació de les activitats pràctiques a partir de l'elaboració de treballs/memòries, exposicions orals i eines d'e-learning de la Universitat (entre el 20% i el 80%).

Avaluació basada en la participació i grau d'implicació de l'alumne en el procés d'ensenyament-aprenentatge, tenint en compte l'assistència regular a les activitats presencials previstes i la resolució de qüestions i problemes proposats periòdicament. (entre el 0 i el 20%).

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