FICHA IDENTIFICATIVA

DATOS DE LA ASIGNATURA

Código: 36429
Nombre: Procesado del lenguaje natural
Ciclo: Grado
Créditos ECTS: 6
Curso académico: 2025-26

TITULACIONES

Titulación	Centro	Curso	Periodo
1400 - Grado en Ingeniería Informática	Escola Tècnica Superior d'Enginyeria	4	Segundo cuatrimeste
1406 - Grado en Ciencia de Datos	Escola Tècnica Superior d'Enginyeria	3	Segundo cuatrimeste

MATERIAS

Titulación	Materia	Carácter
1400 - Grado en Ingeniería Informática	Materia Optativa	OPTATIVA
1406 - Grado en Ciencia de Datos	Aprendizaje automático y minería de datos	OBLIGATORIA

COORDINACIÓN

PELLICER VALERO OSCAR JOSE

RESUMEN

El Procesado del Lenguaje Natural (PLN) es una disciplina en la intersección de la inteligencia artificial, la informática y la lingüística, cuyo objetivo es dotar a las máquinas de la capacidad de comprender, interpretar y generar lenguaje humano. Esta asignatura ofrece un recorrido completo por el campo, partiendo de los fundamentos teóricos y prácticos hasta llegar a las técnicas más avanzadas que definen el estado del arte.

Se abordarán las fases clave de un proyecto de PLN, incluyendo la adquisición, limpieza y preprocesado de texto. Se estudiarán las técnicas fundamentales de representación de texto, desde los modelos clásicos como Bag of Words y TF-IDF hasta los embeddings semánticos que capturan el significado contextual.

El curso aborda también las arquitecturas de Deep Learning que han revolucionado el campo, con un análisis del mecanismo de atención y la arquitectura Transformer. A partir de aquí, se explorará en el funcionamiento y aplicación de los Large Language Models, como los de la familia GPT. Se estudiarán técnicas avanzadas como el fine-tuning eficiente (QLoRA), la ingeniería de prompts, y la construcción de

sistemas complejos como los de Generación Aumentada por Recuperación (RAG).

A través de sesiones teóricas y prácticas, el alumnado adquirirá las competencias necesarias para diseñar e implementar soluciones a problemas reales como la clasificación de textos, la extracción de información, los sistemas de pregunta-respuesta y los agentes conversacionales.

CONOCIMIENTOS PREVIOS

RELACIÓN CON OTRAS ASIGNATURAS DE LA MISMA TITULACIÓN

No se han especificado restricciones de matrícula con otras asignaturas del plan de estudios.

OTROS TIPOS DE REQUISITOS

Para un adecuado seguimiento de la asignatura, se recomienda poseer los siguientes conocimientos:
- Nivel intermedio de programación en Python
- Fundamentos de Machine Learning: Comprensión de los conceptos básicos de aprendizaje supervisado y no supervisado (clasificación, regresión, clustering) y de las métricas de evaluación de modelos.
- Conocimientos básicos de álgebra lineal y probabilidad: Familiaridad con vectores, matrices y conceptos de probabilidad, necesarios para comprender el funcionamiento de los modelos.
- Experiencia previa con librerías del ecosistema de Ciencia de Datos en Python, como pandas, numpy y scikit-learn.

COMPETENCIAS / RESULTADOS DE APRENDIZAJE

-

(CB5) Que los estudiantes hayan desarrollado aquellas habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía.

(CE03) Capacidad para resolver problemas de clasificación, modelización, segmentación y predicción a partir de un conjunto de datos.

(CE07) Capacidad para modelar la dependencia entre una variable respuesta y varias variables explicativas, en conjuntos de datos complejos, mediante técnicas de aprendizaje máquina, interpretando los resultados obtenidos.

(CG06) Capacidad de acceso y gestión de la información en diferentes formatos para su posterior análisis con el fin de obtener conocimiento a partir de datos.

(CT04) Ser responsables de su propio desarrollo profesional y de su especialización, aplicando los conocimientos adquiridos en la identificación de salidas profesionales y yacimientos de empleo.

C1 - Capacidad para conocer los fundamentos, paradigmas y técnicas propias de los sistemas inteligentes y analizar, diseñar y construir sistemas, servicios y aplicaciones informáticas que utilicen dichas técnicas

en cualquier ámbito de aplicación.

C2 - Capacidad para adquirir, obtener, formalizar y representar el conocimiento humano en una forma computable para la resolución de problemas mediante un sistema informático en cualquier ámbito de aplicación, particularmente los relacionados con aspectos de computación, percepción y actuación en ambientes o entornos inteligentes.

C3 - Capacidad para conocer y desarrollar técnicas de aprendizaje computacional y diseñar e implementar aplicaciones y sistemas que las utilicen, incluyendo las dedicadas a extracción automática de información y conocimiento a partir de grandes volúmenes de datos.

DESCRIPCIÓN DE CONTENIDOS

1. Fundamentos y herramientas del procesado del lenguaje natural

Esta unidad introductoria establece las bases conceptuales y prácticas de la asignatura. Se explorará la definición y la historia del Procesado del Lenguaje Natural (PLN), analizando la complejidad inherente del lenguaje humano y sus componentes, desde la morfología hasta la pragmática. A nivel práctico, se abordará el manejo de texto en Python, incluyendo el uso avanzado de cadenas de texto y el poder de las expresiones regulares para la manipulación y búsqueda de patrones. Finalmente, se introducirán técnicas para la adquisición de corpus textuales, como el web scraping, y se presentará el ecosistema de librerías y frameworks que se utilizarán a lo largo del curso, sentando las bases para el ciclo de vida completo de un proyecto de PLN.

2. Ingeniería de texto y análisis lingüístico

El objetivo de esta unidad es aprender a transformar texto crudo y no estructurado en datos limpios y enriquecidos, listos para ser procesados por algoritmos. Se estudiarán en profundidad las técnicas de limpieza y normalización, como la conversión a minúsculas y el tratamiento de caracteres especiales. Se abordará la tarea fundamental de la segmentación (tokenization), tanto a nivel de palabra como de sub-palabra, crucial para los modelos modernos. El contenido se adentra después en el análisis lingüístico computacional, cubriendo el análisis morfológico para la reducción de palabras a su raíz o lema (stemming y lemmatization), el etiquetado gramatical (Part-of-Speech tagging) para identificar la función de cada palabra, el análisis de dependencias sintácticas y el Reconocimiento de Entidades Nombradas (NER).

3. Representación vectorial del texto

Esta unidad se centra en el paso crítico de convertir el texto en representaciones numéricas que una máquina pueda entender, explorando la evolución desde modelos léxicos a semánticos. Se comenzará con las representaciones sparse (dispersas) clásicas, como el modelo Bag-of-Words (BoW) y su refinamiento mediante la ponderación TF-IDF, entendiendo sus fortalezas y limitaciones. A continuación, se producirá un salto conceptual hacia las representaciones densas con los Word Embeddings. Se analizarán en detalle los algoritmos fundacionales como Word2Vec, GloVe y FastText, que revolucionaron el campo al ser capaces de capturar relaciones semánticas y contextuales en un espacio vectorial continuo, permitiendo realizar operaciones algebraicas con el significado de las palabras.

4. Modelado y aplicaciones con vectores de texto

Una vez que el texto ha sido convertido en vectores significativos, esta unidad explora cómo utilizar dichas representaciones para resolver problemas concretos mediante algoritmos de Machine Learning. Se abordará la clasificación de documentos, una de las aplicaciones más comunes del PLN, con casos de uso como el análisis de sentimiento. Se profundizará en la minería de texto no supervisada a través del modelado de tópicos (Topic Modeling), utilizando técnicas como LSI y LDA para descubrir los temas latentes en grandes colecciones de documentos. Finalmente, se estudiarán los fundamentos de los sistemas de recuperación de información (Information Retrieval), desde los algoritmos de ranking clásicos como BM25 hasta los enfoques modernos de búsqueda semántica basados en la similitud de embeddings.

5. Arquitecturas de Deep Learning y modelos contextuales

Esta unidad marca la transición hacia las arquitecturas de redes neuronales profundas que definen el estado del arte actual. Se presentará el contexto histórico con las Redes Neuronales Recurrentes (RNN, LSTM) y Convolucionales (CNN) para el tratamiento de secuencias. El foco principal recaerá en la arquitectura Transformer, desgranando su componente más innovador: el mecanismo de atención (self-attention). Se analizará cómo esta arquitectura permite procesar el texto de forma no secuencial, capturando relaciones a larga distancia. Como aplicación principal de estas ideas, se estudiará en profundidad el modelo BERT (Bidirectional Encoder Representations from Transformers).

6. Grandes Modelos de Lenguaje (LLMs) y aplicaciones generativas

Construyendo sobre la base de la arquitectura Transformer, esta unidad se sumerge por completo en el paradigma de los Grandes Modelos de Lenguaje (LLMs). Se analizará la arquitectura decoder-only característica de la familia de modelos GPT y se detallará su ciclo de vida: desde el pre-entrenamiento a escala masiva hasta las fases de ajuste fino mediante instruction tuning y aprendizaje por refuerzo con retroalimentación humana (RLHF). A nivel práctico, se explorarán las técnicas para interactuar y adaptar estos modelos, como la ingeniería de prompts (prompt engineering), la creación de sistemas de Generación Aumentada por Recuperación (RAG) para basar las respuestas en conocimiento externo, y las metodologías de fine-tuning eficiente como LoRA y QLoRA. Finalmente, se aplicarán estos conocimientos al desarrollo de aplicaciones generativas avanzadas como la traducción automática, el resumen de texto abstractivo y los agentes conversacionales.

7. Laboratorio de Procesado del Lenguaje Natural
Este bloque práctico está diseñado para aplicar de manera tangible los conceptos teóricos desarrollados a lo largo del curso. Las prácticas planificadas son:
Práctica 0: Fundamentos de Python para PLN (no presencial)
Práctica 1: Manipulación de texto con expresiones regulares
Práctica 2: Adquisición de datos textuales: Web Scraping
Práctica 3: Análisis y preprocesado lingüístico
Práctica 4: Modelado y clasificación de documentos
Práctica 5: Aplicaciones con grandes modelos de lenguaje (LLMs)
Práctica 6: Construcción de sistemas con LLMs

VOLUMEN DE TRABAJO (HORAS)

ACTIVIDADES PRESENCIALES

Actividad	Horas
Teoría	30,00
Prácticas en aula	10,00
Laboratorio	20,00
Total horas	**60,00**

ACTIVIDADES NO PRESENCIALES

Actividad	Horas
Asistencia a otras actividades	0,00
Elaboración de trabajos individuales o en grupo	20,00
Estudio y trabajo autónomo	15,00
Preparación de clases	25,00
Preparación de actividades de evaluación	10,00
Resolución de casos prácticos	20,00
Total horas	**90,00**

METODOLOGÍA DOCENTE

Las clases combinarán el contenido teórico con el práctico

MD1 - Actividades teóricas. Desarrollo expositivo de la materia con la participación del alumnado en la resolución de cuestiones puntuales. Realización de cuestionarios individuales de evaluación.

En las actividades teóricas de carácter presencial se desarrollarán los temas de la asignatura proporcionando una visión global e integradora, analizando con mayor detalle los aspectos clave y de mayor complejidad, fomentando, en todo momento, la participación del alumnado (CB05, CT05).

MD2 - Actividades prácticas. Aprendizaje mediante resolución de problemas, ejercicios y casos de estudio a través de los cuales se adquieren competencias sobre los diferentes aspectos teóricos de la materia. (CB05, CG06, CE03, CE07)

Las actividades teóricas se complementan con prácticas de laboratorio con el objetivo de poner en uso los conceptos básicos y ampliarlos con el conocimiento y la experiencia que se vayan adquiriendo durante la realización de los trabajos propuestos.

MD4 -Trabajos en laboratorio y/o aula ordenador. Aprendizaje mediante la realización de actividades guiadas desarrolladas de forma individual o en grupos reducidos y llevadas a cabo en laboratorios y/o aulas de ordenador. (CB05, CG06, CT04, CE03, CE07)

Además de las actividades presenciales, los estudiantes deberán realizar tareas personales (fuera del aula) sobre: cuestiones y problemas, así como la preparación de clases y exámenes (estudio). Estas tareas se realizarán principalmente de manera individual, con el fin de potenciar el trabajo autónomo, pero adicionalmente se incluirán trabajos, especialmente la preparación y resolución de prácticas de laboratorio, que requieran la participación de pequeños grupos de estudiantes (2-3) para fomentar la capacidad de integración en grupos de trabajo.

Se utilizará la plataforma de e-learning (Aula Virtual) de la Universitat de València como soporte de comunicación con el alumnado. A través de ella se tendrá acceso al material didáctico utilizado en clase y

los guiones de las prácticas de laboratorio, así como los problemas y ejercicios a resolver.

EVALUACIÓN

La evaluación del aprendizaje de los conocimientos y competencias conseguidas por el estudiantado se hará de forma continuada a lo largo del curso, y constará de los siguientes bloques:

- SE1 - Prueba objetiva, consistente en un examen que consta tanto de cuestiones teórico-prácticas como de problemas (evaluación de competencias CB05, CT05, CE03, CE07) (50%) (Nota: Todos los porcentajes están referidos a la nota final)

- SE1-1 (40%) Examen de teoría-problemas

- SE1-2 (10%) Examen de laboratorio

- SE2 - Evaluación de las prácticas de laboratorio a partir de la elaboración de trabajos/memorias y/o exposiciones orales (evaluación de competencias CB05, CG06, CT04, CE03, CE07) (35%)

- SE2-1 (20%) Realización de un mini proyecto consistente en el desarrollo de una aplicación completa de PLN para la clasificación de textos

- SE2-2 (15%) Asistencia y evaluación de las sesiones de laboratorio (Actividad NO RECUPERABLE)

- SE3 - Evaluación continua de cada alumno. (15%)

- SE3-1 (15%) Resolución de cuestiones y problemas propuestos (evaluación de competencias CB05, CG06, CE03). (Actividad NO RECUPERABLE)

La nota final de la asignatura se calculará como la media ponderada de cada uno de los apartados anteriores, de acuerdo con el siguiente criterio: SE-1 (50%), SE-2 (35%), SE-3 (15%).

Consideraciones particulares sobre la evaluación:

- Para aprobar la asignatura, es necesario obtener una calificación mínima de 5 (sobre 10) en cada uno de los apartados de evaluación SE1-1, SE1-2 y SE2-1.

- Las actividades SE2-2 y SE3-1 no son recuperables.

En cualquier caso, el sistema de evaluación se regirá por lo establecido en el Reglamento de Evaluación y Calificación de la Universidad de Valencia para Grados y Másteres (https://webges.uv.es/uvTaeWeb/MuestraInformacionEdictoPublicoFrontAction.do?accion=inicio&idEdictoSeleccionado=5639)

La copia o plagio manifiesto de cualquier actividad que forma parte de la evaluación supondrá la imposibilidad de superar la asignatura, sometiéndose seguidamente a los procedimientos disciplinarios oportunos indicados en el PROTOCOLO DE ACTUACIÓN ANTE PRÁCTICAS FRAUDULENTAS EN LA UNIVERSITAT DE VALÈNCIA (ACGUV 123/2020).

BIBLIOGRAFÍA

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