UNIVERSIDAD DE MURCIA

1. Identificación

1.1. De la asignatura

Curso Académico	2024/2025
Titulación	GRADO EN CIENCIA E INGENIERÍA DE DATOS
Nombre de la asignatura	PROCESAMIENTO DE LENGUAJE NATURAL ESCRITO
Código	6600
Curso	TERCERO
Carácter	OBLIGATORIA
Número de grupos	1
Créditos ECTS	6.0
Estimación del volumen de trabajo	150.0
Organización temporal	2º Cuatrimestre
Idiomas en que se imparte	Español

1.2. Del profesorado: Equipo docente

VALENCIA GARCIA, RAFAEL

Docente: **GRUPO 1**

Coordinación de los grupos: **GRUPO 1**

Coordinador de la asignatura

Categoría
CATEDRATICOS DE UNIVERSIDAD

Área
LENGUAJES Y SISTEMAS INFORMÁTICOS

Departamento
INFORMÁTICA Y SISTEMAS

Correo electrónico / Página web / Tutoría electrónica
valencia@um.es http://webs.um.es/valencia Tutoría electrónica: **Sí**

Teléfono, horario y lugar de atención al alumnado

Duración:	Día:	Horario:	Lugar:
C2	Miércoles	10:00-13:00	868888522, Facultad de Informática B1.2.037

Observaciones:
El despacho del profesor es el 2.29 que está en el pasillo central de la segunda planta de la Facultad de Informática. Las tutorías se atenderán por videoconferencia o presencialmente. Por favor, envíen un mensaje privado y se realizará la tutoría en este horario o en otro adecuado.

Duración:	Día:	Horario:	Lugar:
C1	Lunes	10:00-13:00	868888522, Facultad de Informática B1.2.037

Observaciones:
El despacho del profesor es el 2.29 que está en el pasillo central de la segunda planta de la Facultad de Informática. Las tutorías se atenderán por videoconferencia o presencialmente. Por favor, envíen un mensaje privado y se realizará la tutoría en este horario o en otro adecuado.

BERNAL BELTRAN, TOMAS

Docente: **GRUPO 1**

Coordinación de los grupos:

Categoría
CONTRATADO/A PREDOCTORAL (FPU INVES-UM)

Área
No consta

Departamento
No consta

Correo electrónico / Página web / Tutoría electrónica
tomas.bernalb@um.es Tutoría electrónica: **No**

Teléfono, horario y lugar de atención al alumnado

PAN, RONGHAO

Docente: **GRUPO 1**

Coordinación de los grupos:

Categoría
INVESTIGADOR/A LICENCIADO/A

Área
No consta

Departamento
INFORMÁTICA Y SISTEMAS

Correo electrónico / Página web / Tutoría electrónica
ronghao.pan@um.es Tutoría electrónica: **No**

Teléfono, horario y lugar de atención al alumnado

2. Presentación

El objetivo principal de esta asignatura es el de abordar las tecnologías para el Procesamiento del Lenguaje Natural (PLN) diferenciando entre los distintos niveles de procesamiento escrito. Se estudiarán los conceptos de aproximaciones basadas en conocimiento y corpus, los distintos niveles de procesamiento del texto, las representaciones del texto basadas en modelos estadísticos, lingüísticos y de embeddings, los modelos del lenguaje, las arquitecturas de deep learning actuales como Transformers y los modelos del Lenguaje de Gran Escala (LLM).

A nivel técnico, se estudiarán distintas librerías, frameworks. A nivel de ingeniería del software, se estudiará como se pueden abordar distintas tareas de PLN tales como la clasificación de texto, resumen automático, la extracción de entidades, la extracción de tópicos, los sistemas tipo pregunta y respuesta o los chatbots. En este sentido, se analizarán distintos tipos de representación de texto para cada tarea del PLN y se estudiarán las métricas principales para la evaluación del rendimiento y efectividad.

3. Condiciones de acceso a la asignatura

3.1. Incompatibilidades

No constan

3.2. Requisitos

No constan

3.3. Recomendaciones

Se recomienda haber superado las asignaturas Machine learning y haber al menos cursado las asignaturas Machine learning II y Recuperación de información.

4. Competencias

4.1. Competencias básicas

- CB3: Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética
- CB5: Que los estudiantes hayan desarrollado aquellas habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía

4.2. Competencias de la titulación

- CG1: Conocer y aplicar la analítica de datos y técnicas estadísticas apropiadas para descubrir nuevas relaciones en los datos y realizar aportaciones a procesos de las organizaciones, así como apoyar en la toma de decisiones.

- CG4: Capacidad para aplicar los métodos generales de la ciencia e ingeniería de datos en los tipos de datos de dominios específicos, así como en la presentación de los datos, el modelado de datos y procesos, los roles organizacionales y las relaciones entre estos.
- CE23: Aplicar técnicas de procesamiento de lenguaje natural escrito y machine learning, para desplegar modelos para el análisis y la predicción de datos.

4.3. Competencias transversales y de materia

No constan

5. Contenidos

5.1. Teoría

Tema 1: Introducción al Procesamiento del Lenguaje Natural.

- Métodos basados en conocimiento vs métodos basados en corpus.
- Niveles de procesamiento.

Tema 2: Procesamiento léxico morfológico.

- Procesamiento básico del texto.
- Tokenización, stemming, lematización, expresiones regulares.
- POS Tagging.

Tema 3: Procesamiento sintáctico y semántico.

- Análisis sintáctico superficial y análisis de dependencias.
- Análisis semántico. Semántica léxica, distribucional, desambiguación.

Tema 4: Representaciones de texto

- Modelos del lenguaje.
- Modelo de espacio vectorial.
- N-gramas.
- Embeddings.
- Deep learning en PLN:
- Transformers y LLM.
- Zero and Few Shot Learning en PLN.

Tema 5: Tareas y aplicaciones de PLN.

- Clasificación de texto.
- Análisis de sentimientos y emociones..
- Extracción de información.
- Extracción de entidades (NER), extracción de relaciones, expresiones temporales, modelado de tópicos.
- Sistemas pregunta-respuesta. Chatbots.
- Métricas para la evaluación de sistemas de PLN.

5.2. Prácticas

- ### Práctica 1: P1. Aplicaciones prácticas de Procesamiento del lenguaje natural escrito

  Sesiones de prácticas que incluyen la realización de ejercicios sencillos sobre distintas aplicaciones y tecnologías de procesamiento del lenguaje natural relacionados con todos los temas de la asignatura como preprocesamiento de texto, análisis léxico, morfológico, sintáctico, clasificación de texto, extracción de información, reconocimiento de entidades o sistemas de pregunta-respuesta, uso de LLM, etc.

  Estas sesiones se realizan fundamentalmente en los laboratorios de prácticas para aprender contenidos prácticos sobre el uso de librerías y frameworks actuales para el PLN.

  Algunas de estas sesiones prácticas tendrán ejercicios para resolver en casa.

  Relacionado con:
  - Tema 1: Introducción al Procesamiento del Lenguaje Natural.
  - Tema 2: Procesamiento léxico morfológico.
  - Tema 3: Procesamiento sintáctico y semántico.
  - Tema 4: Representaciones de texto
  - Tema 5: Tareas y aplicaciones de PLN.

- ### Práctica 2: P2. Desarrollo de una pequeña aplicación de procesamiento del lenguaje natural

  Se realizará el desarrollo de un proyecto basado en tecnologías de PLN alrededor del caso de estudio de la asignatura como la clasificación de texto o de perfilado de autores. De esta segunda práctica, además del software desarrollado, se realizará un informe técnico que será evaluado para la parte práctica de la asignatura.

  Relacionado con:
  - Tema 2: Procesamiento léxico morfológico.
  - Tema 3: Procesamiento sintáctico y semántico.
  - Tema 4: Representaciones de texto
  - Tema 5: Tareas y aplicaciones de PLN.

6. Actividades Formativas

Actividad Formativa	Metodología	Horas	Presencialidad
AF1: Exposición teórica: Exposición de los contenidos teóricos de la asignatura por parte del profesor. También se contemplan las sesiones informativas sobre el desarrollo del trabajo de fin de grado o prácticas externas.	MD1. Actividades de clase expositiva: Exposición teórica, clase magistral, proyección, dirigida al grupo completo de estudiantes, con independencia de que su contenido sea teórico o práctico/aplicado. Junto a la exposición de conocimientos, en las clases se plantean cuestiones, se aclaran dudas, se realizan ejemplificaciones, se establecen relaciones con las diferentes actividades prácticas que se realizan y se orienta la búsqueda de información.	20.0	100.0
AF2: Seminarios y actividades de aula: Exposición, análisis y debate dentro del contexto de aplicaciones específicas de contenidos teóricos, así como planteamiento y resolución de ejercicios y casos prácticos en el aula, tanto al grupo completo como en grupos reducidos. También se contemplan conferencias, debates y seminarios temáticos.	MD1. Actividades de clase expositiva: Exposición teórica, clase magistral, proyección, dirigida al grupo completo de estudiantes, con independencia de que su contenido sea teórico o práctico/aplicado. Junto a la exposición de conocimientos, en las clases se plantean cuestiones, se aclaran dudas, se realizan ejemplificaciones, se establecen relaciones con las diferentes actividades prácticas que se realizan y se orienta la búsqueda de información.	5.0	100.0
AF3: Prácticas de laboratorio: Ejercicios y resolución de problemas, aprendizaje orientado a proyectos, estudio de casos, exposición y discusión de trabajos, simulaciones y/o prácticas con ordenadores, generalmente desarrolladas en grupos reducidos.	MD4. Actividades prácticas de laboratorio: Actividades de los estudiantes en aulas de informática o específicas, realizadas en grupos reducidos o individualmente, dirigidas al uso y conocimiento de la materia de estudio, supervisadas por el profesor.	28.0	100.0
AF4: Trabajo autónomo del alumno: Estudio y preparación de contenidos teórico-prácticos, trabajo individual consistente en lecturas, búsquedas de información, sistematización de contenidos, elaboración de informes o estudio para la elaboración de casos entre otras actividades.		90.0	0.0
AF5: Tutorías formativas y resolución de dudas: Asistencia individualizada -tutorías individuales- o en grupo -tutorías colectivas- a los estudiantes por parte del profesor.	MD5. Tutorías: Sesiones programadas individuales o en grupo de orientación, revisión o apoyo a los estudiantes por parte del profesor con independencia de que los contenidos sean teóricos o prácticos.	3.0	100.0
AF6: Evaluación: Pruebas individuales, ya sean escritas, orales o con medios informáticos, donde el estudiante demostrará los conocimientos teóricos y prácticos adquiridos durante las actividades formativas asociadas a la enseñanza de la materia.		4.0	100.0
	Totales	150,00	

7. Horario de la asignatura

https://www.um.es/web/estudios/grados/ciencia-ingenieria-datos/2024-25#horarios

8. Sistemas de Evaluación

Identificador	Denominación del instrumento de evaluación	Criterios de Valoración	Ponderación
SE1	Exámenes individuales: Pruebas objetivas, de desarrollo, de respuesta corta, de ejecución de tareas o de escala de actitudes, realizadas por los estudiantes para mostrar los conocimientos teóricos y prácticos adquiridos ya sea de forma oral, escrita o utilizando un ordenador.		40.0
SE3	Procedimientos de observación del trabajo del estudiante: Registros de participación, de realización de actividades, cumplimiento de plazos, participación en foros, informes de seguimiento del trabajo fin de grado y registros sobre el desarrollo de las prácticas externas.		0.0
SE4	Informe técnico. En este instrumento incluimos los resultados de actividades prácticas, o de laboratorio, junto con sus memorias descriptivas y posibles resúmenes del estado del arte sobre temas concretos. La opción de realizar entrevistas personales o presentaciones de los trabajos realizados también entran en esta categoría.		60.0

9. Fechas de exámenes

https://www.um.es/web/estudios/grados/ciencia-ingenieria-datos/2024-25#examenes

10. Resultados del Aprendizaje

- Conocer las características principales de los sistemas de procesamiento del lenguaje natural escrito.
- Aplicar las técnicas de procesamiento del lenguaje natural escrito a partir de fuentes de datos no estructurados más usadas en ciencia e ingeniería de datos.
- Conocer las soluciones computaciones y las posibles plataformas de análisis de texto.
- Elegir las tecnologías adecuadas para desarrollar, estructurar e instrumentar máquinas, experimentos, procesos y sistemas.
- Modelar la aplicación de análisis de texto escrito para desarrollar mejores instrumentos, máquinas, experimentos, procesos y sistemas adecuados.

11. Bibliografía

11. Bibliografía

Bibliografía básica

- Transparencias, apuntes y ejemplos prácticos de la asignatura.
- Dan Jurafsky, James H. Martin (2024). Speech and Language Processing. Stanford University. Disponible online: https://web.stanford.edu/~jurafsky/slp3/
- Tunstall, Lewis, Leandro Von Werra, and Thomas Wolf. Natural language processing with transformers. " O'Reilly Media, Inc.", 2022.

Bibliografía complementaria

- CS224N: Natural Language Processing with Deep Learning. Curso disponible online: https://web.stanford.edu/class/cs224n/
- Bird, Steven, Ewan Klein, and Edward Loper. Natural language processing with Python: analyzing text with the natural language toolkit. " O'Reilly Media, Inc.", 2009.
- Zhai, ChengXiang, and Sean Massung. Text data management and analysis: a practical introduction to information retrieval and text mining. Association for Computing Machinery and Morgan & Claypool, 2016

12. Observaciones

USO DE LA IA: Todos los recursos y materiales no originales que se utilicen en los ejercicios evaluables, incluyendo herramientas de Inteligencia Artificial, ayuda de compañeros, recursos de internet, libros, artículos, etc. deberán referenciarse claramente en el código fuente y la documentación asociada a los ejercicios prácticos entregados.

Los criterios para establecer la nota que aparecerá en el acta son los siguientes:

- Si el alumno no realiza ninguna entrega de prácticas ni se presenta a ninguna evaluación de prueba, su calificación será "No Presentado".
- Si el alumno supera algún instrumento de evaluación y no entrega o no se presenta a los otros, su calificación será "No Presentado".
- Si el alumno no supera un instrumento de evaluación y no presenta o no supera el otro, su calificación será "Suspenso" con la nota del instrumento no superado.
- En otro caso (es decir, se superan los dos intrumentos), su calificación será resultado de aplicar los pesos establecidos para cada instrumento.
- En el caso de plagio, copia o utilización de medios fraudulentos en las diferentes pruebas, se aplicará el artículo 22 del Reglamento Evaluación de Estudiantes (REVA) de la Universidad de Murcia, que conllevará suspender la prueba y, en su caso, podrá ser objeto de sanción previa apertura de expediente académico.

Esta asignatura no se encuentra vinculada de forma directa con los Objetivos de Desarrollo Sostenible.

NECESIDADES EDUCATIVAS ESPECIALES

Aquellos estudiantes con discapacidad o necesidades educativas especiales podrán dirigirse al Servicio de Atención a la Diversidad y Voluntariado (ADYV - https://www.um.es/adyv) para recibir orientación sobre un mejor aprovechamiento de su

proceso formativo y, en su caso, la adopción de medidas de equiparación y de mejora para la inclusión, en virtud de la Resolución Rectoral R-358/2016. El tratamiento de la información sobre este alumnado, en cumplimiento con la LOPD, es de estricta confidencialidad.

REGLAMENTO DE EVALUACIÓN DE ESTUDIANTES

El artículo 8.6 del Reglamento de Evaluación de Estudiantes (REVA) prevé que "salvo en el caso de actividades definidas como obligatorias en la guía docente, si el o la estudiante no puede seguir el proceso de evaluación continua por circunstancias sobrevenidas debidamente justificadas, tendrá derecho a realizar una prueba global".

Se recuerda asimismo que el artículo 22.1 del Reglamento de Evaluación de Estudiantes (REVA) estipula que "el o la estudiante que se valga de conductas fraudulentas, incluida la indebida atribución de identidad o autoría, o esté en posesión de medios o instrumentos que faciliten dichas conductas, obtendrá la calificación de cero en el procedimiento de evaluación y, en su caso, podrá ser objeto de sanción, previa apertura de expediente disciplinario".