24-25

MÁSTER UNIVERSITARIO EN
TECNOLOGÍAS DEL LENGUAJE

GUÍA DE ESTUDIO PÚBLICA

MINERÍA DE TEXTOS

CÓDIGO 31070052

UNED

24-25

MINERÍA DE TEXTOS

CÓDIGO 31070052

ÍNDICE

Nombre de la asignatura	MINERÍA DE TEXTOS
Código	31070052
Curso académico	2024/2025
Título en que se imparte	MÁSTER UNIVERSITARIO EN TECNOLOGÍAS DEL LENGUAJE
Tipo	CONTENIDOS
Nº ETCS	6
Horas	150
Periodo	ANUAL
Idiomas en que se imparte	CASTELLANO

PRESENTACIÓN Y CONTEXTUALIZACIÓN

La asignatura "Minería de textos" se imparte en el Máster Universitario en Tecnologías del Lenguaje. Es una asignatura optativa, de carácter anual, con una carga lectiva de 6 ECTS.
Esta asignatura tiene por objetivo estudiar técnicas de Procesamiento de Lenguaje Natural que permiten analizar el contenido textual de los documentos para transformar información no estructurada en datos estructurados, así como caracterizar los documentos, clasificarlos y agruparlos de forma que se pueda extraer la información relevante para distintas aplicaciones. Se presentan tanto técnicas clásicas de análisis de textos, como técnicas avanzadas de aprendizaje automático y profundo aplicadas al contexto de información textual no estructurada.
El análisis del contenido de los documentos es una parte fundamental de las técnicas actuales de Tecnologías del Lenguaje. Las aplicaciones profesionales son muchas y variadas, incluyendo la minería de opiniones, los sistemas de recomendación, el análisis de redes sociales, la extracción de datos en diferentes dominios, como médico, jurídico, turístico, etc.
Las asignaturas más relacionadas con esta son "Fundamentos del procesamiento lingüístico" y "Minería de Datos". En la primera de ellas se estudian problemas y soluciones (modelos y técnicas) básicas en los niveles de análisis morfológico, sintáctico, semántico y pragmático, mientras que la segunda ofrece una visión panorámica de la teoría y conceptos fundamentales utilizados en Minería de Datos.

REQUISITOS Y/O RECOMENDACIONES PARA CURSAR ESTA ASIGNATURA

No hay ningún requisito diferente de los generales de acceso a este programa de posgrado. Aunque esta asignatura puede ser cursada aisladamente, el estudiante se beneficiaría si hubiera cursado previamente o curse en paralelo la asignatura de *Fundamentos del procesamiento lingüístico*.

EQUIPO DOCENTE

Nombre y Apellidos	RAQUEL MARTINEZ UNANUE
Correo Electrónico	raquel@lsi.uned.es
Teléfono	91398-8725
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	LENGUAJES Y SISTEMAS INFORMÁTICOS

Nombre y Apellidos	M. LOURDES ARAUJO SERNA
Correo Electrónico	lurdes@lsi.uned.es
Teléfono	91398-7318
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	LENGUAJES Y SISTEMAS INFORMÁTICOS

HORARIO DE ATENCIÓN AL ESTUDIANTE

La tutorización de los alumnos se llevará a cabo a través de la plataforma de e-Learning por teléfono y por correo electrónico:

•Raquel Martínez (coodinadora)

email: raquel@lsi.uned.es

Tfno: 913988725

Horario guardias: Martes 09:30 a 13.30

•Lourdes Araujo

email: lurdes@lsi.uned.es

Tfno: 913987318

Horario de guardias: Jueves de 10 a 14.00.

Dirección postal: ETSI Informática, 2ª Planta. C/ Juan del Rosal 16, 28040 Madrid.

COMPETENCIAS QUE ADQUIERE EL ESTUDIANTE

COMPETENCIAS

C2 Abstracción, análisis, síntesis y relación de ideas.

C3 Capacidad crítica y de decisión.

C4 Capacidad de estudio y autoaprendizaje

C5 Capacidad creativa y de investigación.

C7 Capacidad de estudio de los sistemas y aproximaciones existentes y para distinguir las aproximaciones más efectivas.

C8 Capacidad para detectar carencias en el estado actual de la ciencia y la tecnología.

C9 Capacidad para proponer nuevas aproximaciones que de solución a las carencias detectadas.

RESULTADOS DE APRENDIZAJE

CONOCIMIENTOS O CONTENIDOS

CO2 Capacidad de comprender y manejar de forma básica los aspectos más importantes relacionados con los lenguajes y sistemas informáticos en general, y, de manera especial, en los siguientes ámbitos: Tecnologías del lenguaje y de acceso a la información en web.

HABILIDADES O DESTREZAS

H4 Capacidad de especificar, diseñar, implementar y evaluar tanto cualitativa como cuantitativamente los modelos y sistemas propuestos.
H5 Capacidad para proponer y llevar a cabo experimentos con la metodología adecuada como para poder extraer conclusiones y determinar nuevas líneas de actuación e investigación.

COMPETENCIAS

C2 Abstracción, análisis, síntesis y relación de ideas.
C3 Capacidad crítica y de decisión.
C4 Capacidad de estudio y autoaprendizaje.
C5 Capacidad creativa y de investigación.
C7 Capacidad de estudio de los sistemas y aproximaciones existentes y para distinguir las aproximaciones más efectivas.
C8 Capacidad para detectar carencias en el estado actual de la ciencia y la tecnología.
C9 Capacidad para proponer nuevas aproximaciones que de solución a las carencias detectadas.

CONTENIDOS

Tema 1. Introducción

Este primer tema es introductorio, motiva al estudio de la asignatura e introduce los conceptos básicos que se desarrollarán a continuación. Material de estudio disponible en el curso virtual.

Tema 2. Preliminares

•Conceptos básicos de Procesamiento de Lenguaje Natural (PLN)
•Evaluación en PLN: Corpus
•Representación de textos

En este tema se revisan aspectos básicos del PLN, en los que se apoyan los siguientes capítulos. Se introduce el preprocesado de los textos, que trata de normalizar la forma de

los textos, incluyendo técnicas como la normalización de las palabras, la distancia de edición o el tokenizado. Se revisan también algunos niveles de análisis de lenguaje, como el morfológico y el sintáctico, y aspectos como la desambiguación del sentido de las palabras. Se presenta también una introducción a la evaluación de tareas en PLN, incluyendo los corpus. Así mismo se incluye una introducción a las técnicas de representación de textos, tanto vectoriales como distribucionales, funciones de pesado y selección de rasgos, que son fundamentales para posteriores aplicaciones de clasificación, clustering o recuperación de información.
Material de estudio disponible en el curso virtual.

Tema 3. Extracción de información

•Reconocimiento de entidades nombradas
•Extracción de relaciones
•Extracción de sucesos
•Identificación de expresiones temporales

En este capítulo se presentan algunas de las principales tareas de la extracción de información en documentos. Concretamente se incluyen técnicas de reconocimiento de entidades nombradas, tanto no supervisadas (diccionarios, reglas) como supervisadas, incluyendo los sistemas de etiquetado que permiten la clasificación de las entidades. Se describen también las principales técnicas de extracción de relaciones entre entidades y de sucesos y así como de identificación de expresiones temporales en textos.
Material de estudio disponible en el curso virtual.

Tema 4. Clustering

•Métodos de clustering
•Medidas de evaluación
•Herramientas

Se trata de un tema introductorio a una particular manera de organización de objetos, el clustering o agrupación automática. En este caso nos referimos al clustering de documentos, por lo que el contenido se particulariza a este tipo concreto de objetos. Se revisan las principales familias de algoritmos de clustering analizando sus características. Por último, se presentan las medidas de evaluación, estudios comparativos entre diferentes tipos de algoritmos y algunas herramientas de clustering de libre distribución.
Material de estudio disponible en el curso virtual.

Tema 5. Clasificación

•Tipos
•Algoritmos de clasificación
•Medidas de Evaluación
•Herramientas

En este capítulo se proporciona una introducción a la clasificación automática de documentos. En este contexto, y dependiendo de si se dispone o no de datos etiquetados para realizar la tarea de aprendizaje, se distingue entre aprendizaje supervisado y semisupervisado. Se describen los diferentes tipos de clasificación automática, así como las principales técnicas tanto en el aprendizaje supervisado como semisupervisado. Por último, se presentan las medidas de evaluación más usadas dentro de los sistemas de clasificación automática de documentos.
Material de estudio disponible en el curso virtual.

METODOLOGÍA

La metodología es la general del programa de postgrado; junto a las actividades y enlaces con fuentes de información externas, existe material didáctico propio preparado por el equipo docente. Se trata de una metodología adaptada a las directrices del EEES, de acuerdo con el documento del IUED. La asignatura no tiene clases presenciales. Los contenidos teóricos se impartirán a distancia, de acuerdo con las normas y estructuras de soporte telemático de la enseñanza en la UNED.
El temario de la asignatura se estructura en cinco temas y ha sido planteado de tal forma que el estudiante pueda introducirse en los contenidos de la asignatura de una manera gradual, adquiriendo los conocimientos necesarios, y con un enfoque basado en la práctica de los mismos. La búsqueda y estudio de referencias bibliográficas forma parte fundamental del curso.
En cada unidad didáctica elaborada por el equipo docente hay una parte de "Planificación y orientaciones" con la siguiente información:

•Introducción general al contenido.
•Objetivos específicos.
•Esquema de los contenidos.
•Orientaciones sobre la forma de llevar a cabo el estudio del tema.
•Temporización recomendada.
•Indicación de si el tema tiene o no asociada una práctica obligatoria.

El estudiante debe en primer lugar leer esta parte de la unidad didáctica. Las actividades de aprendizaje se estructuran en torno al estado del arte en cada una de las materias del curso y a los problemas en los que se van a focalizar las tareas teórico-prácticas que el alumno

deberá realizar.

Las actividades formativas de la asignatura son:

1. **Actividades teóricas interaccionando con equipos docentes, tutores y compañeros.**
   Resolución de dudas de contenido teórico de forma presencial, vía telefónica o en línea sobre la metodología, los contenidos o las actividades a realizar. Intercambio de información a través de un foro virtual.
2. **Actividades prácticas interaccionando con equipos docentes, tutores y compañeros.**
   Resolución de dudas de contenido práctico de forma presencial, vía telefónica o en línea sobre la metodología, los contenidos o las actividades a realizar. Intercambio de información a través de un foro virtual.
3. **Actividades teóricas desempeñadas autónomamente.**
   Lectura reflexiva y crítica de las orientaciones metodológicas de la asignatura. Estudio de los materiales didácticos.
4. **Actividades prácticas desempeñadas.**
   Elaboración de prácticas o tareas obligatorias de forma individual y en su caso la práctica o tarea opcional.

SISTEMA DE EVALUACIÓN

TIPO DE PRIMERA PRUEBA PRESENCIAL

Tipo de examen	No hay prueba presencial

TIPO DE SEGUNDA PRUEBA PRESENCIAL

Tipo de examen2	No hay prueba presencial

CARACTERÍSTICAS DE LA PRUEBA PRESENCIAL Y/O LOS TRABAJOS

Requiere Presencialidad	No

Descripción

Criterios de evaluación

Aquellos estudiantes que deseen obtener una mayor calificación podrán elegir uno de entre los trabajos optativos que se proponen por parte del equipo docente. En estos casos la calificación final dependerá de la calidad del trabajo realizado.

Ponderación de la prueba presencial y/o los trabajos en la nota final	El promedio de las calificaciones obtenidas en las prácticas obligatorias, incrementado si corresponde (hasta un máximo de 2 puntos) por la práctica opcional constituye la nota final de la asignatura.

Fecha aproximada de entrega

Comentarios y observaciones

La práctica o tarea opcional también tiene un plazo de entrega acorde con la temporización de la asignatura.

PRUEBAS DE EVALUACIÓN CONTINUA (PEC)

¿Hay PEC? No

Descripción

Criterios de evaluación

Ponderación de la PEC en la nota final

Fecha aproximada de entrega

Comentarios y observaciones

OTRAS ACTIVIDADES EVALUABLES

¿Hay otra/s actividad/es evaluable/s? No

Descripción

Criterios de evaluación

Ponderación en la nota final

Fecha aproximada de entrega

Comentarios y observaciones

¿CÓMO SE OBTIENE LA NOTA FINAL?

El promedio de las calificaciones obtenidas en las prácticas obligatorias (un máximo de 8 sobre 10), al que se podrá sumar hasta 2 puntos si se ha realizado la práctica opcional y en función de la calidad de ésta.

BIBLIOGRAFÍA BÁSICA

El equipo docente ha elaborado unidades didácticas para todos los temas de la asignatura. Cada unidad didáctica se compone de:

•Planificación y orientaciones del tema.

•Contenidos teórico-prácticos con enlaces a material disponible en la Web, si es pertinente.

•En caso necesario indicaciones de qué capítulos o partes de la bibliografía básica o complementaria se debe consultar.

Además de las unidades didácticas preparadas por el equipo docente, como bibliografía de la asignatura se deberán estudiar capítulos seleccionados de la siguiente referencia:

•Speech and Language Processing (3rd ed. draft online)

Dan Jurafsky and James H. Martin. (2022) https://web.stanford.edu/~jurafsky/slp3/

BIBLIOGRAFÍA COMPLEMENTARIA

La bibliografía y materiales complementarios se específican de cada capítulo en el material elaborado por el equipo docente.

Puede ser útil el siguiente texto para cuestiones prácticas:

Natural Language Processing with Python
Steven Bird, Ewan Klein, Edward Loper
https://www.nltk.org/book/

RECURSOS DE APOYO Y WEBGRAFÍA

La plataforma de e-Learning proporcionará el adecuado interfaz de interacción entre el alumno y sus profesores. A través de ella se podrá impartir y recibir formación, gestionar y compartir documentos, crear y participar en comunidades temáticas, así como realizar proyectos online.
Se ofrecerán las herramientas necesarias para que, tanto el equipo docente como el alumnado, encuentren la manera de compaginar tanto el trabajo individual como el aprendizaje cooperativo.

IGUALDAD DE GÉNERO

En coherencia con el valor asumido de la igualdad de género, todas las denominaciones que en esta Guía hacen referencia a órganos de gobierno unipersonales, de representación, o miembros de la comunidad universitaria y se efectúan en género masculino, cuando no se hayan sustituido por términos genéricos, se entenderán hechas indistintamente en género femenino o masculino, según el sexo del titular que los desempeñe.