24-25

GUÍA DE ESTUDIO PÚBLICA

RECUPERACIÓN DE INFORMACIÓN Y MINERÍA DE DATOS

CÓDIGO 27040111

UNED

24-25

RECUPERACIÓN DE INFORMACIÓN Y MINERÍA DE DATOS

CÓDIGO 27040111

ÍNDICE

Nombre de la asignatura	RECUPERACIÓN DE INFORMACIÓN Y MINERÍA DE DATOS
Código	27040111
Curso académico	2024/2025
Título en que se imparte	MÁSTER UNIVERSITARIO EN HUMANIDADES DIGITALES: MÉTODOS Y BUENAS PRÁCTICAS
Tipo	CONTENIDOS
Nº ETCS	4
Horas	100
Periodo	SEMESTRE 2
Idiomas en que se imparte	CASTELLANO

PRESENTACIÓN Y CONTEXTUALIZACIÓN

La asignatura "Recuperación de información y minería de datos" está orientada al tratamiento automático de datos textuales. Se divide en dos grandes bloques diferenciados: En la primera parte de la asignatura, se aborda la extracción de información estructurada a partir de contenido textual no estructurado, mediante el reconocimiento de patrones, entidades nombradas, relaciones, etc. Gracias a este primer bloque, el estudiante se familiarizará con las técnicas principales de pre-procesamiento de texto, así como con las herramientas más importantes para la recuperación de información.
En la segunda parte de la asignatura se trabaja sobre el tratamiento de dichos datos utilizando diversas técnicas de representación de la información previamente adquirida. Esta representación del conocimiento nos permite realizar diversas tareas de tratamiento textual: se estudiarán tanto técnicas y algoritmos no supervisados orientados a la agrupación y organización (clustering) de documentación textual, como técnicas supervisadas orientadas a la resolución de tareas de clasificación automática de documentos.

REQUISITOS Y/O RECOMENDACIONES PARA CURSAR ESTA ASIGNATURA

Aunque no existen requisitos obligatorios para cursar esta asignatura, es recomendable el conocimiento de alguna lengua extranjera moderna, preferentemente inglés o francés, para poder acceder a un mayor número de fuentes de recursos (datos, artículos, libros…) que no siempre están traducidos al castellano, o se traducen muy posteriormente a su publicación.
La asignatura guarda relación con las siguientes asignaturas obligatorias del máster: "Análisis y gestión de los datos en la Investigación en Humanidades Digitales" y "Competencias Digitales y Programación para Humanistas". De igual manera, las asignaturas optativas del máster más relacionadas son "Estadística Aplicada" y "Bases de Datos y Big Data".
Se requieren conocimientos de informática básica a nivel usuario, así como el estar familiarizado con herramientas de análisis de datos y/o lenguajes de programación (Python, R, etc).

EQUIPO DOCENTE

Nombre y Apellidos	JUAN MANUEL CIGARRAN RECUERO
Correo Electrónico	juanci@lsi.uned.es
Teléfono	91398-9828
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	LENGUAJES Y SISTEMAS INFORMÁTICOS

Nombre y Apellidos	ANDRES DUQUE FERNANDEZ (Coordinador de asignatura)
Correo Electrónico	aduque@lsi.uned.es
Teléfono	91398-6535
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	LENGUAJES Y SISTEMAS INFORMÁTICOS

HORARIO DE ATENCIÓN AL ESTUDIANTE

Las/os estudiantes pueden contactar con las/os profesoras/es para resolver dudas sobre la asignatura en primer lugar a través del foro de la asignatura en el campus virtual correspondiente, a través del correo electrónico o por teléfono en el horario que se indica. Si se desea una entrevista personal, debe concertarse previamente. En todo tipo de comunicación con el profesorado se deberá indicar la asignatura a la que se refiere y utilizar el correo de la UNED.

Dr. Juan Manuel Cigarrán Recuero
Horario de atención:
Jueves de 10 a 14 horas.
Dirección postal:
Dpto. de Lenguajes y Sistemas Informáticos.
E.T.S.I. Informática. UNED.
C/ Juan del Rosal, 16. 2ª planta. Despacho 2.05.
28040 MADRID
Teléfono: 91.398.7620
Correo electrónico: juanci@lsi.uned.es

Dr. Andrés Duque Fernández
Horario de atención:
Jueves de 11 a 13 horas y de 15 a 17 horas.
Dirección postal:
Dpto. de Lenguajes y Sistemas Informáticos.
E.T.S.I. Informática. UNED.
C/ Juan del Rosal, 16. 2ª planta. Despacho 2.13.
28040 MADRID
Teléfono: 91.398.6535
Correo electrónico: aduque@lsi.uned.es

COMPETENCIAS QUE ADQUIERE EL ESTUDIANTE

COMPETENCIAS BÁSICAS Y GENERALES

CG1 - Administrar el trabajo en equipos multidisciplinares dedicados al ámbito de las Humanidades Digitales de forma eficiente, abordando los posibles conflictos de manera constructiva.

CG2 - Conocer e identificar las nuevas técnicas y herramientas digitales para su empleo en la práctica profesional e investigadora en el ámbito de las Humanidades Digitales.

CG3 - Describir y aplicar las tecnologías para la gestión y organización de la información y la documentación en el ámbito de las Humanidades Digitales.

CB6 - Poseer y comprender conocimientos que aporten una base u oportunidad de ser originales en el desarrollo y/o aplicación de ideas, a menudo en un contexto de investigación.

CB7 - Que los estudiantes sepan aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios (o multidisciplinares) relacionados con su área de estudio.

CB10 - Que los estudiantes posean las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

COMPETENCIAS ESPECÍFICAS

CE3 - Analizar y formalizar la información con herramientas digitales en el ámbito de las Humanidades Digitales.

CE4 - Conocer diferentes formas de gestionar el patrimonio digital de interés para las humanidades.

CE8 - Conocer y saber aplicar diferentes técnicas y tipos de representación de datos digitales y del resultado de su análisis, en el ámbito de las Humanidades Digitales.

CE7 - Aplicar las tecnologías digitales en el tratamiento y la preservación de datos de diferente tipología en el ámbito de las Humanidades Digitales.

CE10 - Explotar corpus textuales (estructurados o no estructurados) de interés para las humanidades.

RESULTADOS DE APRENDIZAJE

- Conocer las principales técnicas relacionadas con la recuperación de información textual.
- Conocer los componentes de una arquitectura básica de recuperación de información: pre-procesado y análisis, reconocimiento de entidades, reglas y sistemas estadísticos para el aprendizaje automático, etc.
- Tener criterios para seleccionar las herramientas actuales de recuperación de información más adecuadas, y familiarizarse con algunas de ellas.
- Saber qué se entiende por minería de textos y conocer las principales técnicas y tecnologías implicadas.

•Saber qué es el clustering de textos y sus características y tipos, así como las implementaciones más utilizadas de los diversos algoritmos.
•Conocer diversos tipos de técnicas de aprendizaje automático que se pueden utilizar en la clasificación automática de textos.

Para alcanzar los resultados anteriores, se realizarán las siguientes actividades formativas:

•Estudio de contenidos
•Tutorías online
•Actividades en la plataforma virtual
•Prácticas informáticas a través del aula virtual
•Trabajos (ensayos, ejercicios, actividades prácticas individuales o en equipo)

CONTENIDOS

TEMA 1

RECUPERACION DE INFORMACION

1. Introducción
2. Recuperación de Información y Procesamiento del Lenguaje Natural
3. Arquitectura de los sistemas de recuperación de información
4. Ejemplos de sistemas de recuperación de información

TEMA 2

REPRESENTACIÓN TEXTUAL PARA MINERIA DE DATOS

1. Introducción
2. Modelos de representación vectorial (VSM)
3. Funciones de pesado
4. Selección y reducción de rasgos
5. Modelos semánticos vectoriales y *Word Embeddings*

TEMA 3

TÉCNICAS DE MINERIA DE TEXTOS. CLUSTERING

1. Introducción
2. Métodos de clustering
3. Trabajos comparativos
4. Técnicas y métricas de evaluación
5. Herramientas

TEMA 4

TÉCNICAS DE MINERIA DE TEXTOS. CLASIFICACIÓN

1. Introducción
2. Aprendizaje automático
3. Tipos de clasificación automática
4. Técnicas supervisadas
5. Técnicas semisupervisadas
6. Técnicas y métricas de evaluación

METODOLOGÍA

La materia está planteada para su realización a través de la metodología general de la UNED, en la que se combinan distintos recursos y los medios impresos con los audiovisuales y virtuales. La metodología estará basada en los siguientes elementos:

1. Materiales de estudio: guía de estudio y web; textos obligatorios; materiales audiovisuales; bibliografía, etc.
2. Participación y utilización de las distintas herramientas del Entorno Virtual de Aprendizaje.
3. Tutorías en línea y telefónica: participación en los foros; comunicación e interacción con el profesorado.
4. Evaluación continua y sumativa: actividades prácticas de evaluación continua; pruebas presenciales; ejercicios de autoevaluación.
5. Trabajo individual o en grupo: lectura analítica de cada tema; elaboración de esquemas; realización de las actividades de aprendizaje propuestas.

SISTEMA DE EVALUACIÓN

TIPO DE PRUEBA PRESENCIAL

Tipo de examen: No hay prueba presencial

CARACTERÍSTICAS DE LA PRUEBA PRESENCIAL Y/O LOS TRABAJOS

Requiere Presencialidad: No

Descripción

NO HAY PRUEBA PRESENCIAL

Criterios de evaluación

Ponderación de la prueba presencial y/o los trabajos en la nota final

Fecha aproximada de entrega

Comentarios y observaciones

PRUEBAS DE EVALUACIÓN CONTINUA (PEC)

¿Hay PEC? Si,PEC no presencial

Descripción

En esta asignatura se realiza una evaluación contínua a través de la elaboración de prácticas obligatorias por tema (hasta un 80% de la nota final)

Criterios de evaluación

Para cada tarea se valorará:

Completitud de la tarea.

Corrección de la tarea.

Ponderación de la PEC en la nota final: 80% de la nota final: 20% PEC Práctica Tema 2, 30% PEC Práctica Tema 3 y 30% PEC Práctica Tema 4

Fecha aproximada de entrega: Las fechas aproximadas de entrega de cada tarea se indicarán en el plan de trabajo de la asignatura

Comentarios y observaciones

OTRAS ACTIVIDADES EVALUABLES

¿Hay otra/s actividad/es evaluable/s? Si,no presencial

Descripción

En esta asignatura se realiza una evaluación contínua a través de la realización de test teóricos al final de cada tema (hasta un 20% de la nota final)

Criterios de evaluación

La corrección de las respuestas en cada uno de los tests teóricos.

Ponderación en la nota final: 20% de la nota final: 5% Test teórico tema 1, 5% test teórico tema 2, 5% test teórico tema 3 y 5% test teórico tema 4.

Fecha aproximada de entrega: Las fechas aproximadas de entrega de cada tarea se indicarán en el plan de trabajo de la asignatura

Comentarios y observaciones

¿CÓMO SE OBTIENE LA NOTA FINAL?

La calificación final máxima será de 10 puntos. Para calcular la nota final de la asignatura se sumarán las notas obtenidas en las PECs prácticas obligatorias de cada tema y los Test teóricos de cada tema con los siguientes pesos:

PECs Prácticas obligatorias por tema -- 80%

Test Teóricos obligatorios por tema -- 20%

Para aprobar la asignatura se exigirá una nota final mínima de 5 puntos, siendo obligatoria la realización de todas las PECs Prácticas y todos los Test Teóricos propuestos en la asignatura.

Se abrirán nuevos plazos de entrega para todas las actividades de cara a la convocatoria de Septiembre.

BIBLIOGRAFÍA BÁSICA

La bibliografía básica será proporcionada al estudiante dentro del curso virtual, estará compuesta por materiales teórico-prácticos realizados por el equipo docente.
Gran parte de la bibliografía, así como los recursos proporcionados al estudiante en el curso virtual pueden estar únicamente en inglés, debido a la novedad de algunos de los contenidos propuestos para la asignatura.

BIBLIOGRAFÍA COMPLEMENTARIA

La bibliografía complementaria de la asignatura se puede encontrar en la sección de "Libros electrónicos" de la biblioteca de la UNED, desde donde se tiene acceso a gran cantidad de recursos online, como puede ser "Safari Books" (O`Reilly), que dispone de una herramienta de búsqueda muy potente para acceder a contenidos online.

- Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. *Introduction to Information Retrieval*, Cambridge University Press. 2008.
- Benjamin Bengfort, Rebecca Bilbro, Tony Ojeda. *Applied Text Analysis with Python*, O'Reilly Media. 2018.
- Bhargav Srinivasa-Desikan. *Natural Language Processing and Computational Linguistics*, Packt Publishing. 2018.
- Grant S. Ingersoll, Thomas S. Morton, and Andrew L. Farris. *Taming Text: How to Find, Organize, and Manipulate It*, Manning Publications. 2012.

RECURSOS DE APOYO Y WEBGRAFÍA

Los/as estudiantes dispondrán de los siguientes recursos de apoyo al estudio:

- Guía de la asignatura. Incluye el plan de trabajo y orientaciones para su desarrollo. Esta guía será accesible desde el curso virtual.
- Curso virtual. A través de esta plataforma los/as estudiantes tienen la posibilidad de consultar información de la asignatura, realizar consultas al Equipo Docente a través de los foros correspondientes, consultar e intercambiar información con el resto de los compañeros/as.
- Documentación de la asignatura. El equipo docente publicará recursos adicionales que faciliten o profundicen los contenidos desarrollados en la asignatura, además de los contenidos ya ofrecidos.
- Biblioteca. El estudiante tendrá acceso tanto a las bibliotecas de los Centros Asociados como a la biblioteca de la Sede Central, en ellas podrá encontrar un entorno adecuado para el estudio, así como de distinta bibliografía que podrá serle de utilidad durante el proceso de aprendizaje.

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.