23-24

MÁSTER UNIVERSITARIO EN
INGENIERÍA Y CIENCIA DE DATOS

GUÍA DE ESTUDIO PÚBLICA

MINERÍA DE TEXTOS

CÓDIGO 31110041

UNED

23-24

MINERÍA DE TEXTOS

CÓDIGO 31110041

ÍNDICE

Nombre de la asignatura	MINERÍA DE TEXTOS
Código	31110041
Curso académico	2023/2024
Título en que se imparte	MÁSTER UNIVERSITARIO EN INGENIERÍA Y CIENCIA DE DATOS
Tipo	CONTENIDOS
Nº ETCS	4
Horas	100.0
Periodo	SEMESTRE 1
Idiomas en que se imparte	CASTELLANO

PRESENTACIÓN Y CONTEXTUALIZACIÓN

Esta asignatura tiene por objetivo estudiar técnicas que permiten transformar información textual no estructurada presente en documentos de distintas clases en datos estructurados fáciles de procesar, que contendrán información relevante y permitirán la extracción de conocimiento. Estos procesos tienen su base en las técnicas de procesamiento de lenguaje natural y aprendizaje automático, que permiten identificar y analizar los elementos informativos de los textos.

Contribución al perfil profesional: La minería de textos tiene muchas aplicaciones dentro de la ciencia de datos ya que hay que tener en cuenta que buena parte del volumen de datos que se maneja son datos no estructurados, texto libre. Esta asignatura permitirá capacitar a los estudiantes para la extracción de este tipo de información y su análisis en grandes volúmenes de documentos de diferentes dominios y de diferentes tipos, incluyendo páginas web, redes sociales, informes médicos, etc.

Se trata de una asignatura obligatoria que se imparte en el primer semestre el máster.

A la asignatura le corresponde 4 créticos ECTS, que equivalen a una estimación de 100 horas de trabajo.

Está relacionada con las siguientes asignaturas:

•Programación en entornos de datos

•Aprendizaje Automático I

En Minería de Textos se presentan librerías y arquitecturas de software específicas para el tratamiento de textos, de ahí su relación con la asignatura Programación en Entornos de Datos. Con respecto a Aprendizaje Automático I, la asignatura de Minería de Textos se centra en los algoritmos de aprendizaje automático aplicados al tratamiento de los textos.

Reseña del profesorado:

ARAUJO SERNA, LOURDES:

Desde 1990 ha desarrollado en universidades públicas diversa actividad docente relacionada con los lenguajes de programación y la algoritmia. Desde 1994 hasta la actualidad ha colaborado de forma continua en programas de doctorado de la Universidad Complutense de Madrid y de la UNED. En la actualidad investiga en procesamiento del lenguaje natural, recuperación de información y en la aplicación de estas técnicas a dominios específicos

como el biomédico.
e.mail: lurdes@lsi.uned.es

MARTÍNEZ UNANUE, RAQUEL:

Ha realizado la mayor parte de su actividad docente en el campo de la programación, la algoritmia, la documentación electrónica y la minería de textos. Su actividad investigadora reciente se centra en la minería de textos, especialmente en representación, clustering y clasificación automática de documentos, tanto monolingües como multilingües, aplicada a diversos tipos de textos (páginas web, noticias, microblogs, historia clínica) y dominios, en particular el dominio biomédico.
e.mail: raquel@lsi.uned.es

REQUISITOS Y/O RECOMENDACIONES PARA CURSAR ESTA ASIGNATURA

Esta asignatura debe estudiarse simultáneamente a las asignaturas de:

•Programación en entornos de datos
•Aprendizaje Automático I

ya que se apoya en los conocimientos impartidos en dichas asignaturas.

Se recomienda que los interesados en cursar el Máster tengan un nivel de lectura en inglés suficiente como para entender contenidos técnicos en dicha lengua. Debido a la novedad de algunos de los contenidos propuestos para la asignatura, gran parte de la bibliografía, así como los recursos proporcionados al estudiante en el curso virtual pueden estar únicamente en inglés

EQUIPO DOCENTE

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

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	AGUSTIN DANIEL DELGADO MUÑOZ
Correo Electrónico	agustin.delgado@lsi.uned.es
Teléfono	91398-8652
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	LENGUAJES Y SISTEMAS INFORMÁTICOS

Nombre y Apellidos	ALVARO RODRIGO YUSTE
Correo Electrónico	alvarory@lsi.uned.es
Teléfono	91398-9693
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	LENGUAJES Y SISTEMAS INFORMÁTICOS

HORARIO DE ATENCIÓN AL ESTUDIANTE

La tutorización de los estudiantes tendrá lugar esencialmente a través de los foros de la plataforma.
En caso de necesitar contactar con el Equipo Docente por medios distintos al curso virtual, se utilizará preferentemente el correo electrónico, utilizando el correo de la asignatura: mitex.cd@lsi.uned.es
pudiéndose también contactar con el equipo docente en los siguientes horarios:
Lourdes Araujo
email: mitex.cd@lsi.uned.es
Tfno: 913987318
Horario guardias: Jueves de 10 a 14.00.
ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA (c/ Juan del Rosal 16, Madrid 28040)
Departamento: LENGUAJES Y SISTEMAS INFORMÁTICOS

Raquel Martínez
email: mitex.cd@lsi.uned.es
Tfno: 913988725
Horario guardias: Martes de 09:30 a 13:30.
ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA (c/ Juan del Rosal 16, Madrid 28040)
Departamento: LENGUAJES Y SISTEMAS INFORMÁTICOS

COMPETENCIAS QUE ADQUIERE EL ESTUDIANTE

COMPETENCIAS BÁSICAS
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.
CB8 - Que los estudiantes sean capaces de integrar conocimientos y enfrentarse a la

complejidad de formular juicios a partir de una información que, siendo incompleta o limitada, incluya reflexiones sobre las responsabilidades. sociales y éticas vinculadas a la aplicación de sus conocimientos y juicios.
CB9 - Que los estudiantes sepan comunicar sus conclusiones y los conocimientos y razones últimas que las sustentan a públicos especializados y no especializados de un modo claro y sin ambigüedades.
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 GENERALES

CG1 - Identificar los métodos apropiados para la solución de problemas asociados a la ciencia de datos y la analítica de información
CG3 - Desarrollar sistemas de gestión/almacenamiento/procesamiento de grandes volúmenes de datos de una manera eficiente y segura, teniendo en cuenta las normativas/legislaciones existentes
CG5 - Utilizar las habilidades de científico de datos y/o ingeniero de datos en entornos de trabajo multidisciplinares y ser capaz de distinguir/organizar las diferentes actividades de los roles en dicho entorno

COMPETENCIAS TRANSVERSALES

CT1 - Ser capaz de abordar y desarrollar proyectos innovadores en entornos científicos, tecnológicos y multidisciplinares.
CT2 - Ser capaz de tomar decisiones y formular juicios basados en criterios objetivos (datos experimentales, científicos o de simulación disponibles).

COMPETENCIAS ESPECÍFICAS

CE7 - Conocer y comprender las técnicas de procesamiento del lenguaje natural (NLP) y su aplicación en la extracción de información en textos
CE9 - Identificar y utilizar técnicas de desarrollo de algoritmos de manipulación de datos en entornos de gestión de datos masivos

RESULTADOS DE APRENDIZAJE

Los resultados más relevantes que se pretende alcanzar con el estudio de esta asignatura son los siguientes:

- Identificar los distintos modelos de extracción de información y análisis de textos, así como las herramientas existentes para el procesamiento de textos.
- Aplicar la metodología de evaluación de sistemas de extracción de información en las fases de desarrollo/implantación de proyectos de procesamiento de textos.
- Discriminar y aplicar los procedimientos necesarios para la búsqueda, selección y manejo de recursos (bibliografía, software, etc.) relacionados con la materia.

Adicionalmente, se pretende que el estudiante alcance los siguientes subobjetivos asociados a los resultados de aprendizaje anteriores:

- Saber qué es la clasificación automática de textos, sus características y tipos.

•Saber utilizar las herramientas disponibles de clasificación automática de textos y tener criterios para seleccionar las más adecuadas.
•Saber qué es el clustering de textos, sus características y tipos.
•Saber utilizar las herramientas disponibles de clustering de textos y tener criterios para seleccionar las más adecuadas.
•Conocer diversas aplicaciones de la minería de textos.

CONTENIDOS

Introducción al procesamiento del lenguaje natural.

En este tema se presentan diversas tareas básicas de procesamiento del lenguaje natural, que sirven de base para tratar otros problemas más complejos. Se presentan herramientas prácticas para abordar dichas tareas básicas.

Extracción de información en documentos.

En este tema se estudian técnicas encaminadas a identificar en un documento los datos relevantes para un problema considerado, así como su estructura y relaciones. Concretamente se abordan tareas relacionadas con la identificación de entidades, conceptos y sus relaciones en documentos.

Representación de documentos.

En este capítulo se proporciona una introducción a la representación automática de textos y a los modelos de representación más utilizados en minería de textos.

Clasificación y clustering.

En este capítulo se presentan la clasificación y clustering de documentos, se revisan las principales familias de algoritmos analizando sus características. Por último, se presentan algunas herramientas de libre distribución.

Aplicaciones.

Este capítulo presenta las características principales de algunas aplicaciones de actualidad de las técnicas de minería de textos.

METODOLOGÍA

Esta asignatura ha sido diseñada para la enseñanza a distancia. Por tanto, el sistema de enseñanza-aprendizaje estará basado en gran parte en el estudio independiente o autónomo del estudiante. Para ello, el estudiante contará con diversos materiales que permitirán su trabajo autónomo y la Guía de Estudio de la asignatura, que incluye orientaciones para la realización de las actividades prácticas. Asimismo, mediante la plataforma virtual de la UNED existirá un contacto continuo entre el equipo docente y los/as estudiantes, así como una interrelación entre los propios estudiantes a través de los foros, importantísimo en la enseñanza no presencial.
El estudio de esta asignatura se realizará a través de los materiales que el Equipo Docente publicará en el curso virtual.
La asignatura tiene un carácter eminentemente práctico. Se presentan contenidos fundamentales de campo del Procesamiento del Lenguaje Natural, centrándose en el uso de herramientas para abordar con facilidad problemas prácticos que se presentan al manejar información no estructurada. Estas técnicas son especialmente relevantes en los ámbitos en los que es necesario trabajar con grandes cantidades de información.

Se fomentará el uso de software libre siempre y cuando sea posible para la realización de las actividades y las practicas propuestas.

Los temas van acompañados de prácticas, en algunos casos obligatorias para aprobar la asignatura, que proporcionan al estudiante capacidad para abordar tareas de procesamiento del lenguaje en distintos ámbitos.
Las actividades formativas para el estudio de la asignatura son las siguientes:
•Estudios de contenidos (45 horas)
•Tutorías (8 horas)
•Actividades en la plataforma virtual (2 horas)
•Prácticas informáticas (40 horas)
•Otros trabajos y examen (5 horas)
•Total:100 horas

SISTEMA DE EVALUACIÓN

TIPO DE PRUEBA PRESENCIAL

Tipo de examen	Examen de desarrollo
Preguntas desarrollo	6
Duración del examen	120 (minutos)
Material permitido en el examen	
Ninguno	
Criterios de evaluación	

Normas de valoración del examen:

La nota del examen representa el 60% de la valoración final de la asignatura (el 40% restante corresponde a las prácticas obligatorias).

% del examen sobre la nota final	60
Nota del examen para aprobar sin PEC	
Nota máxima que aporta el examen a la calificación final sin PEC	
Nota mínima en el examen para sumar la PEC	
Comentarios y observaciones	

CARACTERÍSTICAS DE LA PRUEBA PRESENCIAL Y/O LOS TRABAJOS

Requiere Presencialidad	Si

Descripción

La prueba presencial, tanto de febrero, como de septiembre, consta de 6 de cuestiones sobre el temario de la asignatura.

Criterios de evaluación

La nota total del examen debe ser al menos de 3 sobre 6 para compensar con la nota de las prácticas.

En caso de que solo se apruebe el examen o las prácticas en la convocatoria de febrero, se guarda la parte aprobada

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?	No

Descripción

Esta asignatura tiene Prácticas en lugar de PED.

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?	Si,no presencial

Descripción

El trabajo del curso incluye la realización de dos prácticas obligatorias. El objetivo de estas prácticas es ayudar al alumno a la comprensión de los temas tratados, así como hacerle ver su aplicación.

El enunciado de las prácticas estará disponible en el curso virtual de la asignatura.

Criterios de evaluación

Las prácticas son corregidas por el equipo docente. Cada práctica, calificada de 0 a 10 supone un 20% de la nota de la asignatura. Así, la nota asignada podrá incrementar hasta un máximo de 4 puntos (2 por cada práctica) en la nota final de la asignatura. A modo de ejemplo se tendrán las siguientes correspondencias:

Sobresaliente (10) -> 2

Sobresaliente (9) -> 1.8

Notable (7) -> 1.4

Aprobado (5) -> 1

Ponderación en la nota final 40%

Fecha aproximada de entrega

Comentarios y observaciones

Las prácticas tendrán una fecha de entrega especificada en el enunciado y la entrega se realizará en la plataforma virtual.

Las fechas de entrega de las prácticas son aproximadamente en la primera midad de diciembre y a mediados de enero. La fecha exacta se anunciará en el entorno virtual.

En el curso virtual se facilitarán con el material de cada tema ejercicios de autoevaluación.

¿CÓMO SE OBTIENE LA NOTA FINAL?

Las prácticas y el examen deben aprobarse por separado.

La nota del examen representa el 60% de la valoración final de la asignatura y las prácticas el 40% restante.

BIBLIOGRAFÍA BÁSICA

ISBN(13):null
Título:NATURAL LANGUAGE PROCESSING WITH PYTHONnull
Autor/es:
Editorial:sin publicar

ISBN(13):null
Título:SPEECH AND LANGUAGE PROCESSING3ª
Autor/es:
Editorial:sin publicar

La bibliografía básica no incluye algunos contenidos del curso. Por ello, en el entorno virtual de la asignatura se pondrá a disposición de los alumnos material de estudio complementario (presentaciones, artículos, recopilaciones y referencias a otro material disponible en la web).

Los libros propuestos se encuentran en Internet.

BIBLIOGRAFÍA COMPLEMENTARIA

ISBN(13):9789811052088
Título:DEEP LEARNING IN NATURAL LANGUAGE PROCESSING2018
Autor/es:
Editorial:Springer

RECURSOS DE APOYO Y WEBGRAFÍA

La plataforma de e-Learning Alf proporcionará el adecuado interfaz de interacción entre el alumno y sus profesores. Alf es una plataforma de e-Learning y colaboración que permite 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.