

19-20

MÁSTER UNIVERSITARIO EN
INGENIERÍA Y CIENCIA DE DATOS

GUÍA DE ESTUDIO PÚBLICA



MINERÍA DE TEXTOS

CÓDIGO 31110041

Ambito: GUI - La autenticidad, validez e integridad de este documento puede ser verificada mediante el
Código Seguro de Verificación (CSV) en la dirección <https://sede.uned.es/valida/>



90634BB3DA406F5F4B1162A0EB86C84F

uned

19-20

MINERÍA DE TEXTOS

CÓDIGO 31110041

ÍNDICE

PRESENTACIÓN Y CONTEXTUALIZACIÓN
REQUISITOS Y/O RECOMENDACIONES PARA CURSAR ESTA ASIGNATURA
EQUIPO DOCENTE
HORARIO DE ATENCIÓN AL ESTUDIANTE
COMPETENCIAS QUE ADQUIERE EL ESTUDIANTE
RESULTADOS DE APRENDIZAJE
CONTENIDOS
METODOLOGÍA
SISTEMA DE EVALUACIÓN
BIBLIOGRAFÍA BÁSICA
BIBLIOGRAFÍA COMPLEMENTARIA
RECURSOS DE APOYO Y WEBGRAFÍA
ADENDA AL SISTEMA DE EVALUACIÓN CON MOTIVO DE LA PANDEMIA COVID 19

Ámbito: GUI - La autenticidad, validez e integridad de este documento puede ser verificada mediante el "Código Seguro de Verificación (CSV)" en la dirección <https://sede.uned.es/valida/>



90634BB3DA406F5F4B1162A0EB86C84F

Nombre de la asignatura	MINERÍA DE TEXTOS
Código	31110041
Curso académico	2019/2020
Título en que se imparte	MÁSTER UNIVERSITARIO EN INGENIERÍA Y CIENCIA DE DATOS
Tipo	CONTENIDOS
Nº ECTS	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.

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

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

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

Ámbito: GUI - La autenticidad, validez e integridad de este documento puede ser verificada mediante el "Código Seguro de Verificación (CSV)" en la dirección <https://sede.uned.es/valida/>



90634BB3DA406F5F4B1162A0EB86C84F

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
Correo Electrónico
Teléfono
Facultad
Departamento

M. LOURDES ARAUJO SERNA (Coordinador de asignatura)
lurdes@lsi.uned.es
91398-7318
ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
LENGUAJES Y SISTEMAS INFORMÁTICOS

Nombre y Apellidos
Correo Electrónico
Teléfono
Facultad
Departamento

RAQUEL MARTINEZ UNANUE
raquel@lsi.uned.es
ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
LENGUAJES Y SISTEMAS INFORMÁTICOS

HORARIO DE ATENCIÓN AL ESTUDIANTE

Lourdes Araujo (coordinadora)
email: lurdes@lsi.uned.es
Tfno: 913987318
Horario guardias: Jueves de 10 a 14.00.

Raquel Martínez
email: raquel@lsi.uned.es
Tfno: 913988725
Horario guardias: Miércoles de 09:30 a 13.30.

Ámbito: GUI - La autenticidad, validez e integridad de este documento puede ser verificada mediante el "Código Seguro de Verificación (CSV)" en la dirección <https://sede.uned.es/valida/>



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

Ámbito: GUI - La autenticidad, validez e integridad de este documento puede ser verificada mediante el "Código Seguro de Verificación (CSV)" en la dirección <https://sede.uned.es/valida/>



90634BB3DA406F5F4B1162A0EB86C84F

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.

Ámbito: GUI - La autenticidad, validez e integridad de este documento puede ser verificada mediante el "Código Seguro de Verificación (CSV)" en la dirección <https://sede.uned.es/valida/>



90634BB3DA406F5F4B1162A0EB86C84F

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

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)
- Trabajos (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

Ámbito: GUI - La autenticidad, validez e integridad de este documento puede ser verificada mediante el "Código Seguro de Verificación (CSV)" en la dirección <https://sede.uned.es/valida/>



90634BB3DA406F5F4B1162A0EB86C84F

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).

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

% 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 consta de 6 cuestiones sobre el temario de la asignatura.

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? 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

Ámbito: GUI - La autenticidad, validez e integridad de este documento puede ser verificada mediante el "Código Seguro de Verificación (CSV)" en la dirección <https://sede.uned.es/validar>



90634BB3DA406F5F4B1162A0EB86C84F

Las prácticas son corregidas por el equipo docente. La nota asignada podrá incrementar hasta un máximo de 4 puntos (2 por cada práctica) en la nota final de la asignatura de acuerdo con el siguiente baremo:

Sobresaliente -> + 2

Notable -> + 1.5

Aprobado -> + 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.

¿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):

Título:NATURAL LANGUAGE PROCESSING WITH PYTHON

Autor/es:Steven Bird ; Edward Loper ; Ewan Klein ;

Editorial:sin publicar

ISBN(13):

Título:SPEECH AND LANGUAGE PROCESSING (3ª)

Autor/es:James H. Martin ; Jurafsky, Dan ;

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

Ámbito: GUI - La autenticidad, validez e integridad de este documento puede ser verificada mediante el "Código Seguro de Verificación (CSV)" en la dirección <https://sede.uned.es/valida/90634BB3DA406F5F4B1162A0EB86C84F>



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.

ADENDA AL SISTEMA DE EVALUACIÓN CON MOTIVO DE LA PANDEMIA COVID 19

<https://app.uned.es/evacaldos/asignatura/adendasig/31110041>

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.

Ámbito: GUI - La autenticidad, validez e integridad de este documento puede ser verificada mediante el "Código Seguro de Verificación (CSV)" en la dirección <https://sede.uned.es/valida/>



90634BB3DA406F5F4B1162A0EB86C84F