24-25

GUÍA DE ESTUDIO PÚBLICA

DESCUBRIMIENTO DE INFORMACIÓN EN TEXTOS

CÓDIGO 31101254

UNED

24-25

DESCUBRIMIENTO DE INFORMACIÓN EN TEXTOS

CÓDIGO 31101254

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
IGUALDAD DE GÉNERO

Nombre de la asignatura	DESCUBRIMIENTO DE INFORMACIÓN EN TEXTOS
Código	31101254
Curso académico	2024/2025
Título en que se imparte	MÁSTER UNIVERSITARIO EN INVESTIGACIÓN EN INTELIGENCIA ARTIFICIAL
Tipo	CONTENIDOS
Nº ETCS	6
Horas	150
Periodo	ANUAL
Idiomas en que se imparte	CASTELLANO

PRESENTACIÓN Y CONTEXTUALIZACIÓN

La asignatura "Descubrimiento de información en textos" se imparte en el Máster Universitario en Investigación en Inteligencia Artificial. 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 de los documentos, así como caracterizarlos, 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 de documentos es una parte fundamental de las técnicas actuales de Tecnologías del Lenguaje ya que permite extraer datos específicos de grandes volúmenes de textos, además la clasificación y agrupamiento de documentos son fundamentales para encontrar información relevante para una necesidad de información específica. En cuanto a 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, aportando un enfoque orientado a su uso, independientemente de la implementación particular.

REQUISITOS Y/O RECOMENDACIONES PARA CURSAR ESTA ASIGNATURA

No hay ningún requisito diferente de los generales de acceso a este programa de posgrado orientado a la investigación. 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	M LOURDES ARAUJO SERNA
Correo Electrónico	lurdes@lsi.uned.es
Teléfono	
Facultad	
Departamento	

Nombre y Apellidos	RAQUEL MARTINEZ UNANUE (Coordinador/a de asignatura)
Correo Electrónico	raquel@lsi.uned.es
Teléfono	
Facultad	
Departamento	

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 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

CG2 - 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.

CG1 - 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.

CG3 - 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.

CG4 - 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

CE2 - Conocer un conjunto de métodos y técnicas tanto simbólicas como conexionistas y probabilistas, para resolver problemas propios de la Inteligencia Artificial.

CE3 - Conocer los procedimientos específicos de aplicación de estos métodos a un conjunto relevante de dominio (educación, medicina, ingeniería, sistemas de seguridad y vigilancia, etc.), que representan las áreas más activas de investigación en IA.

CE1 - Conocer los fundamentos de la Inteligencia Artificial y las fronteras actuales en investigación.

RESULTADOS DE APRENDIZAJE

El objetivo del curso es proporcionar al alumno una visión global de las técnicas y tecnologías involucradas en el descubrimiento de información en textos.

El aprendizaje está diseñado para permitir que el alumno adquiera una serie de *destrezas y competencias* que se enumeran a continuación:

- Saber lo que es un corpus y conocer los criterios por los que se clasifican, los tipos de anotaciones más comunes y los estándares utilizados.
- Conocer los modelos de representación comúnmente utilizados, así como los métodos de selección y reducción del número de rasgos.
- Saber distinguir los diversos niveles de información lingüística que se pueden utilizar en la representación de textos y las notaciones para su descripción.
- Saber qué se entiende por minería de textos y conocer las principales técnicas y tecnologías implicadas.
- Saber qué es la clasificación automática de textos y sus características y tipos.
- Conocer diversos tipos de técnicas de aprendizaje automático que se pueden utilizar en la clasificación automática de textos.

•Conocer los modelos estadísticos más utilizados en el procesamiento del lenguaje.
•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 y sus características y tipos.
•Conocer diversos tipos de algoritmos de clustering.
•Saber utilizar las herramientas disponibles de clustering de textos y tener criterios para seleccionar las más adecuadas.
•Conocer algoritmos de etiquetado léxico y análisis sintáctico.

CONTENIDOS

Tema 1. Introducción.

1. Definiciones preliminares.
2. Interés y aplicaciones.

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. Evaluación en PLN: Corpus y otros aspectos

1. Corpus
2. Campañas de evaluación
3. Medidas de evaluación
4. Baselines
5. Análisis de errores

Este capítulo introduce una serie de aspectos relacionados con la evaluación de las aplicaciones en PLN. Se proporciona una introducción a las compilaciones de textos o corpus utilizados en el procesamiento del lenguaje natural. Estos textos pueden estar o no anotados con información lingüística. Se describen distintos tipos de corpus y anotaciones, y se presentan ejemplos. Se describen así mismo las campañas de evaluación, presentando algunos ejemplos. Finalmente se introducen distintos aspectos de la evaluación como medidas utilidad frecuentemente, los resultados base de una evalaución o baselines y el análisis de errores.

Tema 3. Estándares de anotaciones.

1. Introducción
2. Lenguajes de anotaciones. XML
1. Generalidades
2. Componentes de un documento XML
3. Modelado de datos
4. Fundamentos de las DTD
5. Corrección de un documento XML
3. Estándares de anotaciones en XML
1. TEI
2. XCES
4. Anotaciones *stand-off* en XML.
5. JSON

En este capítulo se proporciona una introducción sobre los tipos de anotaciones más comunes en corpus textuales. Este tipo de anotaciones facilitan diversas tareas relacionadas con la minería de textos. El lenguaje más común utilizado hoy en día para anotar corpus es XML. Se proporciona una introducción que podrán saltarse aquellos alumnos que ya dispongan de conocimientos al respecto.

A continuación se presentan dos de los estándares XML más utilizados por la comunidad científica así como por profesionales. Uno es muy general, TEI, y el otro, XCES, más específico de las anotaciones con información lingüística. Ambos se utilizan en Ingeniería Lingüística y en aplicaciones de Procesamiento de Lenguaje Natural.

Seguidamente, se introduce la arquitectura de las anotaciones stand-off en XML, que permite superar algunas de las limitaciones intrínsecas de XML y facilitar el procesamiento de textos anotados.

Por último, se presenta JSON un formato de intercambio de datos independiente del lenguaje muy extendido en la actualidad.

Tema 4. Modelos para la caracterización de textos

1. Preprocesado de texto
2. Análisis morfológico, lematización y stemming
3. Etiquetado Léxico (POS tagging)
4. Chunking
5. Análisis sintáctico (Parsing)
6. Desambiguación del sentido de las palabras
7. Detección de entidades nombradas

Este capítulo introduce diversos procesos básicos realizados habitualmente para caracterizar textos, y que tienen aplicación en otros procesos como la extracción de información, o la clasificación de textos. Comenzamos por introducir diversas técnicas de preprocesado de textos (eliminación de palabras vacías, normalización de la forma de las palabras, etc.) que se aplican como paso previo a otras técnicas, tratando de uniformizar la forma del texto. Se presentan después distintos tipos de análisis de los contenidos de los textos, comenzado por el morfológico que se aplica a palabras, hasta el sintáctico que se aplica a oraciones. Hacemos particular énfasis en los modelos de asignación de categoría léxica (POS tagging) y de análisis sintáctico. Se introducen también otras técnicas importantes para la caracterización de textos como son la desambiguación del sentido de las palabras y el reconocimiento de entidades nombradas.

Tema 5. Representación de textos: Modelos y funciones de pesado y de reducción de rasgos.

1. Introducción.
2. Modelos de representación vectorial.
3. Funciones de pesado.
4. Funciones de selección y reducción de rasgos

En este capítulo se proporciona una introducción a la representación automática de textos. Además del clásico modelo de espacio vectorial, en este tema se introducen otros modelos y en especial un modelo de representación que ha tomado una especial relevancia en los últimos años, los conocidos como word embeddings. En su forma más básica en su aplicación al Procesamiento del Lenguajes Natural se trata de representaciones vectoriales de términos, en lo que se conoce como representaciones distribuidas, y que se han desarrollado en parte gracias al éxito del deep learning (aprendizaje profundo).
Además, se presentan funciones de ponderación empleadas para calcular la importancia o relevancia de una cadena en el contenido de un texto. Estas funciones pueden emplear parámetros diferentes según los casos; desde la frecuencia de aparición en el documento o en la colección, hasta probabilidades condicionadas en problemas de clasificación automática.
Se introducen también aspectos relacionados con la selección de rasgos (conjunto de cadenas con el que se va a representar) como elementos de transformación de una información que inicialmente es de carácter cualitativo.

Tema 6. Técnicas de minería de textos. Clustering.

1. Introducción
2. Métodos de clustering
3. Trabajos comparativos
4. Medidas de evaluación
5. 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 estudios comparativos entre diferentes tipos de algoritmos y algunas herramientas de clustering de libre distribución.

Tema 7. Técnicas de minería de textos. Clasificación automática.

1. Introducción.
2. Tipos de clasificación automática.
3. Algoritmos de clasificación.
4. Evaluación.

En este capítulo se proporciona una introducción a la clasificación automática de documentos dentro del *Aprendizaje Automático.* 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 funciones de evaluación más usadas dentro de los sistemas de clasificación automática de documentos.

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 siete temas y ha sido planteado de tal forma que el alumno 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 específícos.

•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. Como se trata de un máster orientado a la investigación, 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 teorico-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

No hay prueba presencial y las prácticas no requieren presencialidad.

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 no se realiza una prueba presencial, la evaluación se realiza mediante evaluación continua a partir de las siguientes pruebas:
Las prácticas obligatorias a lo largo del curso.
La práctica opcional para subir nota, una vez que se han realizado las prácticas obligatorias.
Aquellos alumnos que deseen obtener una mayor calificación en la convocatoria de junio, podrán elegir uno de entre los trabajos optativos que se irán proponiendo. En estos casos la calificación final dependerá de la calidad del trabajo realizado.
Las tareas obligatorias se deberán entregar en los plazos que se vayan indicando. La no entrega de las tareas en el plazo previsto supondrá suspender la asignatura en la convocatoria de junio. El trabajo optativo para subir nota también tendrá una fecha límite de entrega. Habrá otro plazo de entrega de tareas para la convocatoria de septiembre.

Criterios de evaluación

En la asignatura se realizarán cuatro prácticas obligatorias cuya entrega es un requisito imprescindible para aprobar la asignatura. Dos de las prácticas se corresponderán con los contenidos de los temas 2, 3 y 4, y las otras dos prácticas con los contenidos de los temas 5, 6 y 7.

La realización correcta de todas las prácticas obligatorias asegura una nota de APROBADO, que podría llegar hasta NOTABLE (8) dependiendo de la calidad de las soluciones en su conjunto.

Aquellos estudiantes que deseen obtener una mayor calificación podrán elegir uno de entre los trabajos optativos que se proponen por parte del equipo docente, normalmente se ofertan 2 o 3 trabajos optativos. En estos casos la calificación final dependerá de la calidad del trabajo realizado.

Ponderación de la PEC en la nota final	El promedio de las calificaciones obtenidas en las prácticas obligatorias y en su caso en la práctica opcional constituye la nota final de la asignatura.
Fecha aproximada de entrega	
Comentarios y observaciones	

Las prácticas obligatorias tienen un plazo de entrega fijo, que suele ser de unas tres semanas después de haber finalizado el tema correspondiente, de acuerdo con la temporización de la asignatura y los periodos vacacionales. Esta temporización permite al estudiante suficiente margen de tiempo para poder organizar su trabajo de acuerdo con sus circunstancias personales.

Los estudiantes que no entreguen las tareas en el plazo establecido para la convocatoria de junio tendrán otro plazo de entrega en la convocatoria de septiembre.

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

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 indica qué capítulos o partes de la bibliografía básica o complementaria se debe consultar.

Como bibliografía de la asignatura se deberán estudiar capítulos seleccionados de las siguientes referencias:

•Speech and Language Processing (3rd ed. draft online)
Dan Jurafsky and James H. Martin. (2022)

•Gordon, A.D. Classification. 2nd Edition. Chapman &Hall/CRC, 1999.

•Mitchell, T. Machine Learning. McGraw Hill, 1997. (Nuevos capítulos creados en 2006 y disponibles en https://www.cs.cmu.edu/%7Etom/mlbook.html)

•S. Weiss; N. Indurkhya; T. Zhang; F. Damerau. Text Mining: Predictive Methods for Analyzing Unstructured Information, 2004.

BIBLIOGRAFÍA COMPLEMENTARIA

La bibliografía complementaria es específica de cada capítulo e incluye partes de libros y artículos. A modo de muestra se presentan aquí algunos de ellos:

•[Manning et al. 2008] Manning, C.D., Raghavan, p. y Schütze, H. Introduction to Information Retrieval. Cambridge University Press, 2008.

•[Salton 1989] Salton, G. Automatic text processing: the transformation, analysis, and retrieval of information by computer. Addison-Wesley Longman Publishing Co., Inc., 1989.

•[Zhao &Karypis 2002] Zhao, Y. y Karypis, G. Evaluation of Hierarchical Clustering Algorithms for Document Datasets. Data Mining and Knowledge Discovery, pag. 515 524, ACM Press, 2002.

•[Milligan &Cooper 1985] Milligan, G. y Cooper, M. An examination of procedures for determining the number of clusters in a data set . Psychometrika, 50(2), pag. 159 179, 1985.

•[Sebastiani 2002] Sebastiani, F. Machine learning in automated text categorization. ACM Comput. Surv., 34(1), pag. 1 47, 2002.

•[Joachims 1998] Joachims, T. Text categorization with support vector machines: learning with many relevant features. Proceedings of ECML-98, 10th European Conference on Machine Learning, pag. 137 142, 1998.

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.