

24-25

GUÍA DE ESTUDIO PÚBLICA



REPRESENTACIÓN DE TEXTOS EN ESPACIOS VECTORIALES Y PROBABILÍSTICOS

CÓDIGO 31070023

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/>



DC287E8FD0FAA4C3B2847D61B8E277A0



24-25

**REPRESENTACIÓN DE TEXTOS EN
ESPACIOS VECTORIALES Y
PROBABILÍSTICOS
CÓDIGO 31070023**

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

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/>



DC287E8FD0FAA4C3B2847D61B8E277A0

Nombre de la asignatura	REPRESENTACIÓN DE TEXTOS EN ESPACIOS VECTORIALES Y PROBABILÍSTICOS
Código	31070023
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 "Descubrimiento de información en textos" se enmarca dentro del Máster en Tecnologías del Lenguaje impartido por la Escuela Técnica Superior de Ingeniería Informática de la UNED.

Ficha técnica:

- Tipo: Optativa
- Duración: Anual
- Créditos Totales y Horas: 6 / 150
- Horas de estudio teórico: 75
- Horas de trabajo práctico: 75

Reseña del Profesorado:

FRESNO FERNÁNDEZ, VÍCTOR

Víctor Fresno forma parte del grupo NLP&IR de la UNED. Sus líneas de investigación se centran fundamentalmente en el estudio y propuesta de modelos de representación de textos para su procesamiento automático y su aplicación a problemas de Clasificación Automática, Agrupamiento y Recuperación de Información. Realizó una estancia de investigación post-doctoral como Visiting Faculty en la City University of New York (CUNY). Desde el año 2000 hasta la actualidad ha trabajado en el Instituto de Automática industrial (CSIC), la Universidad Rey Juan Carlos (URJC) y la Universidad Nacional de Educación a Distancia (UNED), colaborando en los programas de doctorado de dichas universidades. e.mail: vfresno@lsi.uned.es

AMIGÓ CABRERA, ENRIQUE

Enrique Amigó forma parte del grupo NLP&IR de la UNED. Sus líneas de investigación se centran en: (i) la axiomatización de métricas de evaluación y su conexión con teoría de la medida, (ii) la extensión de la teoría de la información para rasgos continuos en representación de documentos y formalización del concepto de similitud, y más recientemente (iii) la formalización de la sinergia entre composicionalidad y contextualidad en modelos de representación semántica. Sus trabajos cuentan con un total de 2400 citas

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/>



DC287E8FD0FAA4C3B2847D61B8E277A0

según Google Scholar. Entre otros méritos, destacan el premio Google Faculty Research Award 2012 junto con los investigadores Julio Gonzalo y Stefano Mizzaro, y la organización del congreso internacional SIGIR 2022 en Madrid.

e.mail: enrique@lsi.uned.es

REQUISITOS Y/O RECOMENDACIONES PARA CURSAR ESTA ASIGNATURA

Conocimientos previos recomendables:

- Diseño e implementación de sistemas informáticos.
- Lectura fluida del inglés.
- Fundamentos matemáticos de la informática.

Esta asignatura puede ser cursada aisladamente, aunque el estudiante se beneficiaría si hubiera cursado previamente o cursara en paralelo la asignatura de *Fundamentos del Procesamiento Lingüístico*, y las asignaturas de Fundamentos Matemáticos de la Informática y Estadística impartidas en el primer ciclo de la titulación de Informática de la UNED, o asignaturas equivalentes en otras universidades.

EQUIPO DOCENTE

Nombre y Apellidos
Correo Electrónico
Teléfono
Facultad
Departamento

ENRIQUE AMIGO CABRERA
enrique@lsi.uned.es
91398-8651
ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
LENGUAJES Y SISTEMAS INFORMÁTICOS

Nombre y Apellidos
Correo Electrónico
Teléfono
Facultad
Departamento

VICTOR DIEGO FRESNO FERNANDEZ (Coordinador de asignatura)
vfresno@lsi.uned.es
91398-8217
ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
LENGUAJES Y SISTEMAS INFORMÁTICOS

Nombre y Apellidos
Correo Electrónico
Teléfono
Facultad
Departamento

ALEJANDRO BENITO SANTOS
al.benito@lsi.uned.es
ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
LENGUAJES Y SISTEMAS INFORMÁTICOS

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/>



DC287E8FD0FAA4C3B2847D61B8E277A0

HORARIO DE ATENCIÓN AL ESTUDIANTE

La tutorización de los alumnos se llevará a cabo a través de la plataforma online de la UNED, por teléfono y por correo electrónico:

•Enrique Amigó

email: enrique@lsi.uned.es

Tfno: 913988651

Horario guardias: Jueves de 15:00 a 19:00

•Víctor Fresno

email: vfresno@lsi.uned.es

Tfno: 913988217

Horario guardias: Martes y Miércoles de 11:30 a 13:30

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

COMPETENCIAS QUE ADQUIERE EL ESTUDIANTE

COMPETENCIAS

C1 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.

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.

C6 Habilidades sociales para el trabajo en equipo

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

CO1 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.

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.

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/>



DC287E8FD0FAA4C3B2847D61B8E277A0

HABILIDADES O DESTREZAS

H1 Capacidad de aplicar los conocimientos adquiridos y su capacidad de resolución de problemas en entornos nuevos o poco conocidos dentro de contextos más amplios relacionados con su área de estudio.

H2 Capacidad 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.

H3 Poseer las habilidades de aprendizaje que les permitan continuar estudiando de un modo que habrá de ser en gran medida autodirigido o autónomo.

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

C1 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.

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.

C6 Habilidades sociales para el trabajo en equipo.

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 a la representación de textos.

Visión global de los principales paradigmas de representación de textos, ventajas y debilidades de cada uno de ellos. Presentación de la estructura del curso.

Tema 2: Modelo de Espacio Vectorial (Vector Space Model)

1. Fundamento teórico: Principio de independencia
2. Funciones de selección y pesado de rasgos.
3. Técnicas de reducción de dimensionalidad.



Tema 3: Modelos de lenguaje (Language Models)

1. Nociones fundamentales de un modelo de lenguaje.
2. Estimación de un modelo de lenguaje.
3. Perplejidad como medida de evaluación de modelos de lenguaje y ajuste entre textos y modelos.

Tema 4: Modelos de representación neuronal: una visión de conjunto.

1. Dimensiones sobre las que categorizar los modelos
2. Modelos superficiales.
3. Modelos recurrentes.
4. Modelos recursivos.
5. Modelos convolucionales.
6. Mecanismos de atención.

Tema 5: Semántica vectorial: word embeddings.

1. Conceptos básicos: Noción de semántica distribucional.
2. De la representación de documentos basada en bolsa de palabras a la representación léxica.
3. Teoría de la información y representación distribucional: producto de vectores vs. información mutua.
4. Modelo Skip-grams
5. Propiedades semánticas de la representación semántico distribucional.

Tema 6: Modelos de lenguaje neuronales.

1. Conceptos básicos de las redes neuronales: La neurona, redes de avance (Feedforward Networks), múltiples capas, funciones de coste.
2. Características generales de un modelo de lenguaje neuronal.
3. Embedding de palabras.
4. Modelado de lenguaje mediante una red.

Tema 7: Arquitecturas de aprendizaje profundo para procesamiento de secuencias de palabras.

1. Limitaciones de las redes de avance para el modelado de lenguaje.
2. Redes recurrentes.



3. Redes recurrentes bidireccionales.
4. Long Short-term Memory (LSTM).
5. Mecanismos de atención

Tema 8: Geometría de la semántica distribucional.

1. Antecedentes: Semántica y filosofía del lenguaje.
2. Embedding. Repaso de los mecanismos de representación basados en modelos de lenguaje. Medidas de cantidad de información: Relación entre la noción de cantidad de información y los modelos de lenguaje. Cantidad de información en embedding estáticos y contextuales.
3. Orientación semántica, pragmática, representación lógica y significado.
4. Contextualidad y composicionalidad.
5. Isometría e isotropía: Correspondencia entre proximidad en el espacio de embeddings versus espacio semántico.

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 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 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. 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 teórico-prácticas que el alumno deberá realizar.

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/>



DC287E8FD0FAA4C3B2847D61B8E277A0

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.

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

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/>



DC287E8FD0FAA4C3B2847D61B8E277A0

Se recuerda al estudiantado que no está permitido el uso de herramientas de Inteligencia Artificial Generativa (IAG) para la elaboración de trabajos académicos derivados del desarrollo de la asignatura, salvo indicación expresa en contra por parte del Equipo Docente.

En cualquier caso, sobre las posibilidades y límites en el uso de este tipo de herramientas en la UNED, puede consultarse la "Guía de uso de las herramientas de Inteligencia Artificial Generativa para el estudiantado" elaborada por el Vicerrectorado de Innovación Educativa y disponible en este sitio web:

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 tareas obligatorias teórico-prácticas. **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. Habrá otro plazo de entrega de tareas para la convocatoria de septiembre.**

Criterios de evaluación

Los temas del programa de la asignatura a partir del Tema 2 tienen asociada una tarea teórico-práctica obligatoria cuya entrega es un requisito imprescindible para aprobar la asignatura. Cada tarea se calificará con una nota de 0 a 10, y tendrán la misma ponderación dentro del curso.

Ponderación de la PEC en la nota final

El promedio de las calificaciones obtenidas en las tareas teórico-prácticas constituye la nota final de la asignatura.

Fecha aproximada de entrega

Comentarios y observaciones

Las tareas asociadas a cada tema tienen un plazo de entrega fijo, 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.

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>



DC287E8FD0FAA4C3B2847D61B8E277A0

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 tareas teórico-prácticas constituye la nota final de la asignatura, siempre que todas ellas tengan una calificación mínima de 5.

BIBLIOGRAFÍA BÁSICA**Bibliografía básica:**

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

Bibliografía complementaria:

- Como bibliografía complementaria se aportarán referencias dentro del curso virtual.

El equipo docente ha elaborado Unidades Didácticas para todos los temas de la asignatura.

Cada unidad didáctica se compone de documentos 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.

BIBLIOGRAFÍA COMPLEMENTARIA**RECURSOS DE APOYO Y WEBGRAFÍA**

Los 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.

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/>



DC287E8FD0FAA4C3B2847D61B8E277A0

- **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.

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/>



DC287E8FD0FAA4C3B2847D61B8E277A0