23-24

MÁSTER UNIVERSITARIO EN
TECNOLOGÍAS DEL LENGUAJE

GUÍA DE ESTUDIO PÚBLICA

REPRESENTACIÓN DE TEXTOS EN ESPACIOS VECTORIALES Y PROBABILÍSTICOS

CÓDIGO 31070023

UNED

23-24

REPRESENTACIÓN DE TEXTOS EN ESPACIOS VECTORIALES Y PROBABILÍSTICOS

CÓDIGO 31070023

ÍNDICE

Nombre de la asignatura	REPRESENTACIÓN DE TEXTOS EN ESPACIOS VECTORIALES Y PROBABILÍSTICOS
Código	31070023
Curso académico	2023/2024
Título en que se imparte	MÁSTER UNIVERSITARIO EN TECNOLOGÍAS DEL LENGUAJE
Tipo	CONTENIDOS
Nº ETCS	6
Horas	150.0
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 totat de 2400 citas

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 congreo 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 Lingüistico,* 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	ENRIQUE AMIGO CABRERA
Correo Electrónico	enrique@lsi.uned.es
Teléfono	91398-8651
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	LENGUAJES Y SISTEMAS INFORMÁTICOS

Nombre y Apellidos	VICTOR DIEGO FRESNO FERNANDEZ (Coordinador de asignatura)
Correo Electrónico	vfresno@lsi.uned.es
Teléfono	91398-8217
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	LENGUAJES Y SISTEMAS INFORMÁTICOS

Nombre y Apellidos	ALEJANDRO BENITO SANTOS
Correo Electrónico	al.benito@lsi.uned.es
Teléfono	
Facultad	ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
Departamento	LENGUAJES Y SISTEMAS INFORMÁTICOS

HORARIO DE ATENCIÓN AL ESTUDIANTE

La tutorización de los alumnos se llevará a cabo a través de la plataforma de e-Learning Alf, 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
COMPETENCIAS GENERALES
CPG1 - Adquirir capacidad de abstracción, análisis, síntesis y relación de ideas.
CPG2 - Adquirir capacidad crítica y de decisión
CPG3 - Adquirir capacidad de estudio y autoaprendizaje
CPG4 - Adquirir capacidad creativa y de investigación
CPG5 - Adquirir habilidades sociales para el trabajo en equipo
COMPETENCIAS ESPECÍFICAS
CE1 - Adquirir 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.
CE3 - Adquirir capacidad de estudio de los sistemas y aproximaciones existentes para distinguir las aproximaciones más efectivas.
CE4 - Adquirir capacidad para detectar carencias en el estado actual de la ciencia y la tecnología.
CE5 - Adquirir capacidad para proponer nuevas aproximaciones que den solución a las carencias detectadas.
CE6 - Adquirir capacidad de especificar, diseñar, implementary evaluar tanto cualitativa como cuantitativamente los modelos y sistemas propuestos.
CE7 - Adquirir 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

RESULTADOS DE APRENDIZAJE

El objetivo del curso es proporcionar al alumno una visión global sobre los modelos y técnicas de representación de textos dentro de los diferentes paradigmas de las Tecnologías de la Lengua.
El aprendizaje está diseñado para permitir que el alumno adquiera una serie de *destrezas y competencias* que se enumeran a continuación:

- Adquirir una visión general de los diferentes modelos de representación de textos existentes en la literatura, adquiriendo una serie de destrezas y competencias que se enumeran a continuación:
- Entender la representación textos dentro del Modelo de Espacio Vectorial, proyectando un texto en un espacio multidimensional definido a partir de un vocabulario.
- Conocer los modelos de lenguaje. Entender la interpretación del lenguaje como las posibles secuencias de términos y su distribución de probabilidad.
- Conocer los modelos basados en arquitecturas neuronales, en donde la representación del texto se basa en el estado interno de la red.
- Hábito de lectura de artículos científicos y capacidad para buscar información que complete el material propuesto inicialmente.
- Capacidad de reflexión sobre el material estudiado, necesaria para poder realizar una síntesis de calidad.
- Desarrollar pequeñas aplicaciones para la obtención de diferentes modelos de representación de textos.

Con la superación del curso se espera que el alumno complete todas las competencias generales especificadas en la memoria del máster:
CPG1 - Adquirir capacidad de abstracción, análisis, síntesis y relación de ideas.
CPG2 - Adquirir capacidad crítica y de decisión
CPG3 - Adquirir capacidad de estudio y autoaprendizaje
CPG4 - Adquirir capacidad creativa y de investigación
CPG5 - Adquirir habilidades sociales para el trabajo en equipo

Además de las siguientes competencias específicas:
CE1 - Adquirir 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.
CE3 - Adquirir capacidad de estudio de los sistemas y aproximaciones existentes para distinguir las aproximaciones más efectivas.
CE4 - Adquirir capacidad para detectar carencias en el estado actual de la ciencia y la

tecnología.
CE5 - Adquirir capacidad para proponer nuevas aproximaciones que den solución a las carencias detectadas.
CE6 - Adquirir capacidad de especificar, dise.ar, implementar y evaluar tanto cualitativa como cuantitativamente los modelos y sistemas propuestos.
CE7 - Adquirir 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.

CONTENIDOS

Tema 1. Introducción a la representación de textos.

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. N-gramas.
2. Perplejidad.
3. Técnicas de suavizado.

Tema 4: Modelos basados en Redes Neuronales (Neural Based Representation Models)

1. Semántica vectorial: *word embeddings*.
2. *Compositional sentence embeddings*.
3. Modelos de lenguaje neuronales.
4. Arquitecturas de aprendizaje profundo para procesamiento de secuencias de palabras.
5. Geometría de la semántica distribucional.

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 cuatro 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 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 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.

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.

BIBLIOGRAFÍA BÁSICA

Bibliografía básica:

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

Bibliografía coplementaria:

•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.
- **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.