

26-27

GUÍA DE ESTUDIO PÚBLICA



REPRESENTACIÓN DE TEXTOS EN ESPACIOS VECTORIALES Y PROBABILÍSTICOS

CÓDIGO 31070023

UNED

26-27

**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

Nombre de la asignatura	REPRESENTACIÓN DE TEXTOS EN ESPACIOS VECTORIALES Y PROBABILÍSTICOS
Código	31070023
Curso académico	2026/2027
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. Su línea de investigación se centra fundamentalmente en el estudio y propuesta de modelos de representación de textos para su procesamiento automático y su aplicación a problemas de clasificación automática, clustering y recuperación de información. Actualmente está centrado en la línea de composición semántica distribucional y estimación de similitud dentro de espacios vectoriales y probabilísticos, y en el contexto de los modelos neurales del lenguaje. 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 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 las asignaturas de *Fundamentos del Procesamiento Lingüístico* y *Redes Neuronales para el Procesamiento del Lenguaje Natural*, así como 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/a 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
91398-6484
ESCUELA TÉCN.SUP INGENIERÍA INFORMÁTICA
LENGUAJES Y SISTEMAS INFORMÁTICOS

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 de 11:00 a 15:00

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.

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

BLOQUE I: CONCEPTOS PREVIOS

En este BLOQUE I se introducen los fundamentos teóricos y metodológicos de la representación de textos. Se estudian la evolución histórica de los principales paradigmas, las teorías del significado, el modelo de espacio vectorial y los modelos de lenguaje probabilísticos.

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. Teorías del significado

Este tema introduce las principales teorías del significado y su relación con la representación de textos en espacios vectoriales y probabilísticos. Se abordará cómo distintas concepciones del significado influyen en la forma de modelar y representar las expresiones lingüísticas en procesamiento del lenguaje natural.

El tema permitirá comprender los fundamentos teóricos que subyacen a diferentes modelos de representación, así como sus posibilidades, limitaciones y ámbitos de aplicación.

Tema 3. Modelo de Espacio Vectorial

Este tema presenta el modelo de espacio vectorial como uno de los enfoques fundamentales para la representación de textos en procesamiento del lenguaje natural. Se estudiará cómo los documentos pueden representarse mediante vectores de rasgos y cómo esta representación permite abordar tareas como la recuperación de información, la clasificación o el clustering de documentos.

El tema introducirá los principales componentes del modelo, organizándolos en los siguientes apartados:

1. Fundamentos teóricos del modelo de espacio vectorial:
2. Selección y pesado de rasgos.
3. Técnicas de reducción de dimensionalidad.

Tema 4. Modelos de lenguaje

Este tema introduce los fundamentos de los modelos de lenguaje como mecanismos para estimar la probabilidad de secuencias de palabras. Se estudiará su papel en la representación y el procesamiento de textos, así como su relación con distintos fenómenos lingüísticos y con tareas predictivas como la clasificación o la inferencia sobre textos.

El tema combina una perspectiva probabilística clásica con una visión general de los enfoques neuronales actuales, prestando atención tanto a la estimación de distribuciones de probabilidad como a su evaluación y ajuste.

Los contenidos se organizan en los siguientes apartados:

1. Nociones fundamentales de un modelo de lenguaje
2. Estimación de modelos de lenguaje
3. Evaluación de modelos de lenguaje
4. Técnicas de suavizado

BLOQUE II: SEMÁNTICA DISTRIBUCIONAL EN MODELOS NEURONALES

En este BLOQUE II se aborda la representación semántica mediante embeddings y modelos neuronales. Se estudian los fundamentos de la semántica distribucional, los modelos de lenguaje neuronales, distintas arquitecturas de representación y las propiedades geométricas de los embeddings.

Tema 5. Semántica vectorial: embeddings

Este tema introduce la semántica vectorial y los embeddings léxicos como mecanismos de representación densa del significado de las palabras. Se estudiará la hipótesis distribucional como fundamento lingüístico de estos modelos y se analizará cómo los contextos de aparición permiten construir representaciones semánticas útiles para tareas de procesamiento del lenguaje natural.

El tema conecta los embeddings con los modelos de espacio vectorial y los modelos de lenguaje estudiados previamente, prestando atención a su interpretación geométrica, su relación con la teoría de la información y sus principales ventajas y limitaciones frente a otros enfoques de representación.

Los contenidos se organizan en los siguientes apartados:

1. Fundamentos de la semántica distribucional
2. De la representación documental a la representación léxica
3. Teoría de la información y representación distribucional
4. Modelos de embeddings léxicos
5. Propiedades semánticas de los embeddings

Tema 6. Modelos de lenguaje neuronales

Este tema introduce los fundamentos de los modelos de lenguaje neuronales y su aplicación a la representación y predicción de secuencias lingüísticas. Se estudiarán los principios básicos de las redes neuronales, desde la neurona artificial hasta arquitecturas sencillas de avance, con el fin de comprender cómo estos modelos pueden adaptarse al procesamiento del lenguaje natural.

El tema presta especial atención al papel de los embeddings como representaciones distribuidas de las palabras dentro de modelos predictivos, así como a las diferencias entre los modelos neuronales y los enfoques probabilísticos clásicos estudiados previamente.

Los contenidos se organizan en los siguientes apartados:

1. Fundamentos de redes neuronales
2. Características generales de los modelos de lenguaje neuronales
3. Embeddings de palabras
4. Modelado de lenguaje mediante redes neuronales

Tema 7. Arquitecturas neuronales de representación

Este tema profundiza en las arquitecturas neuronales utilizadas para la representación del lenguaje, analizando su relación con la semántica distribucional y con distintas teorías lingüísticas. El tema permitirá comprender las capacidades y limitaciones de distintas arquitecturas neuronales desde una perspectiva lingüística y representacional. Los contenidos abordados servirán como base conceptual para la práctica final de la asignatura.

Tema 8. Distribución de embeddings en el espacio y capas neuronales: anisotropías

Este tema estudia la distribución geométrica de los embeddings en los espacios vectoriales generados por modelos neuronales profundos. Se analizarán fenómenos como la anisotropía y la forma en que las representaciones evolucionan a través de las distintas capas de una red neuronal.

BLOQUE III: OPERADORES SOBRE EMBEDDINGS

Este BLOQUE III se centra en las operaciones que pueden definirse sobre representaciones vectoriales. Se estudian operadores de cuantificación, composición, similitud y otras transformaciones semánticas, como analogía, implicación, generalización o reconstrucción.

Tema 9. Operador de cuantificación

Este tema analiza la relación entre las representaciones vectoriales y distintas medidas de cantidad de información asociadas a los textos. Se estudiará cómo magnitudes como el *Pointwise Mutual Information* o el *Information Content* pueden vincularse con los embeddings, con el objetivo de mejorar su interpretación y su uso en operaciones semánticas.

Tema 10. Composicionalidad lingüística

Este tema estudia la composicionalidad lingüística en el contexto de las representaciones vectoriales. Se analizará cómo los embeddings de unidades básicas, como palabras o tokens, pueden combinarse para obtener representaciones de unidades lingüísticas más complejas, como sintagmas, oraciones o textos completos.

Tema 11. Operadores de similitud

Este tema aborda el estudio de distintas funciones de similitud entre embeddings, atendiendo tanto a sus propiedades matemáticas como a su interpretación semántica. Se retomarán las funciones de similitud introducidas en el modelo de espacio vectorial y se ampliará su análisis al contexto de las representaciones distribucionales y neuronales.

Tema 12. Otros operadores

Este tema presenta una panorámica de otros operadores semánticos relevantes en el trabajo con embeddings y representaciones vectoriales. Se estudiarán operaciones como la analogía, la implicación textual, la generalización, la especialización y distintos mecanismos de reconstrucción.

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.

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

En relación a 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.

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 tres bloques en los que se estructura el programa de la asignatura tienen asociadas tareas prácticas obligatorias 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. Con la realización de estas prácticas obligatorias se podrá llegar hasta una nota máxima de 7.

Se propondrán ejercicios teórico opcionales, dentro de cada uno de los bloques del curso, que permitirán subir la nota final de la asignatura hasta el 10.

Ponderación de la PEC en la nota final

El promedio de las calificaciones obtenidas en las diferentes tareas teórico/prácticas constituirá la nota final de la asignatura.

Fecha aproximada de entrega

Comentarios y observaciones

Las tareas prácticas asociadas a cada bloque 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, siempre que todas ellas tengan una calificación mínima de 5.

BIBLIOGRAFÍA BÁSICA

ISBN(13): 9780135041963

Título: SPEECH AND LANGUAGE PROCESSING segunda edición

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

Editorial: PEARSON EDUCATION

Bibliografía básica (versión **online** actualizada):

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

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, donde se indican qué capítulos o partes de la bibliografía básica o complementaria se debe consultar.
- Contenidos teórico-prácticos multimedia complementados con enlaces a material disponible en la web, si es pertinente.

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 propios y adicionales que faciliten o profundicen en los contenidos desarrollados en la asignatura.
- **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.