



COMILLAS

UNIVERSIDAD PONTIFICIA

ICAI

ICADE

CIHS

GUÍA DOCENTE

2026 - 2027

FICHA TÉCNICA DE LA ASIGNATURA

Datos de la asignatura	
Nombre completo	Datos Masivos (Big Data)
Código	DCIA-TITLE-521
Título	Graduado o Graduada en Traducción, Interpretación y Tecnologías del Lenguaje por la Universidad Pontificia Comillas [GRIRL 23]
Impartido en	Grado en Traducción, Interpretación y Tecnologías del Lenguaje [Cuarto Curso] Grado en Traducción, Interpretación y Tecnologías del Lenguaje [Tercer Curso]
Créditos	3,0 ECTS
Carácter	Optativa
Departamento / Área	Departamento de Ciencias de la Computación e Inteligencia Artificial

Datos del profesorado	
Profesor	
Nombre	Blanca Hernández Pardo
Departamento / Área	Departamento de Traducción e Interpretación y Comunicación Multilingüe
Despacho	Cantoblanco [D-303]
Correo electrónico	bhpardo@comillas.edu

DATOS ESPECÍFICOS DE LA ASIGNATURA

Contextualización de la asignatura
Aportación al perfil profesional de la titulación
En el perfil del graduado en Traducción, Interpretación y Tecnologías del Lenguaje, esta asignatura aporta la capacidad de trabajar con grandes volúmenes de datos lingüísticos, aplicando herramientas y metodologías del ámbito del Big Data y la ciencia de datos al análisis y procesamiento del lenguaje.
Prerrequisitos
Se recomienda haber cursado Fundamentos de programación y Programación orientada a PLN.

Resultados del proceso de formación y aprendizaje	
Conocimientos o contenidos	
CN 10	Conoce los elementos básicos del desarrollo de la IA aplicada al lenguaje humano, tales como anotación de diverso tipo, ontologías y reconocimiento de voz.
CN 7	Conoce los fundamentos de la programación en su aplicación al procesamiento del lenguaje natural para la mediación interlingüística.



CN 8	Conoce los procedimientos de inteligencia artificial en aspectos centrales como aprendizaje automático, aprendizaje profundo o procesamiento del lenguaje
Habilidades o destrezas	
HB 1	Puede aplicar teorías y conceptos lingüísticos a la comprensión y análisis de la lengua y evaluar la calidad de la comunicación en diferentes contextos.
HB 5	Puede trabajar en equipo con otros profesionales del lenguaje, la comunicación y el PLN.
HB 6	Puede adaptarse a los cambios tecnológicos y a las nuevas tendencias en el campo del multilingüismo, y desarrollar habilidades de aprendizaje autónomo y continuo en dicho ámbito.

BLOQUES TEMÁTICOS Y CONTENIDOS

Contenidos – Bloques Temáticos

0. Presentación de la asignatura y ecosistema de herramientas

- Organización de la asignatura: sesiones, evaluación
- El ecosistema de trabajo: Python y Jupyter Notebook
- Por qué se trabaja con un entorno virtual reproducible
- Mapa general de las unidades de la asignatura

1. Introducción al análisis de datos masivos

- Características de los datos masivos: volumen, variedad y velocidad
- La dimensión crítica y social del «big data»
- De la analítica de negocio a la analítica aplicada al lenguaje
- El flujo CRISP-DM aplicado a contextos lingüísticos y documentales
- Primer contacto práctico con la escala en Python
- Relación entre datos masivos e inteligencia artificial: los sistemas de IA (incluida la IA generativa) se entrenan y dependen de grandes volúmenes de datos; la calidad, representatividad y sesgo de esos datos condiciona su comportamiento

2. Preparación, limpieza y transformación de datos a escala

- Importación de datos y selección de variables
- Estructuración de tablas y detección de inconsistencias
- Transformación y organización de datasets amplios
- Limpieza orientada al análisis lingüístico/documental (Pandas)
- Actividad de investigación docente «descarga cognitiva e IA»: limpieza de una columna de texto (sin IA) y detección de valores atípicos con el método IQR (con apoyo de un asistente de IA generativa, revisando y justificando cada sugerencia), con medición de retención inmediata y diferida

3. Exploración y visualización de datos masivos

- Distribuciones, histogramas, boxplots, scatter plots y heat maps
- Tablas y vistas exploratorias



- Interpretación prudente de patrones en conjuntos de datos amplios
- Herramientas: Pandas, Matplotlib y Seaborn

4. Corpus y colecciones documentales extensas: organización y catalogación a gran escala

- De la colección de archivos al corpus: por qué catalogar antes de analizar
- Organización de archivos y carpetas a escala
- Recorrido de una colección de documentos con Python (os, pathlib)
- Extracción de metadatos y construcción de un catálogo tabular con Pandas

5. Almacenamiento y acceso a datos a gran escala: bases de datos y fuentes externas

- El modelo relacional: tablas, claves y consultas
- Acceso a una base de datos relacional desde Python (SQLite)
- Acceso a fuentes de datos externas a través de una URL
- Integración de datos procedentes de fuentes distintas y exportación a base de datos

6. Automatización y comunicación de resultados a gran escala

- De la tarea puntual a la tarea repetida a escala
- Conversión de un cuaderno en script Python reutilizable
- Funciones que generan informes y exportación a Excel multi-hoja
- Comunicación de resultados mediante gráficos (Matplotlib)
- Buenas prácticas de automatización y reproducibilidad

METODOLOGÍA DOCENTE

Aspectos metodológicos generales de la asignatura

Metodología eminentemente práctica basada en el aprendizaje por proyectos y casos: el alumno trabaja con conjuntos de datos reales, desarrolla y ejecuta flujos de tratamiento y análisis, y reflexiona sobre los resultados. Se emplean entornos accesibles y colaborativos (p. ej. cuadernos en la nube).

RESUMEN HORAS DE TRABAJO DEL ALUMNO

HORAS PRESENCIALES		
AF1 - Lecciones de carácter expositivo - Exposición en que el/a docente explica una serie de nociones de forma clara, estructurada y motivadora con mayor o menor grado de participación activa y colaborativa del alumnado. Incluirá presentaciones dinámicas y la participación reglada o espontánea del estudiantado por medio de actividades diversas.	AF2 - Ejercicios prácticos/resolución de problemas - Ejercicios y resolución de problemas planteados por el/a docente a partir de una breve lectura, un material preparado para la ocasión, o cualquier otro tipo de datos o informaciones que supongan un desafío intelectual para el/ la estudiante.	AF3 - Trabajos individuales/grupales - Trabajos de carácter teórico o práctico, individuales o grupales, que implican la lectura de artículos, revistas, informes de investigación, capítulos de libros, informaciones en Internet, etc. y la redacción de una reflexión personal (de diverso calado y extensión) que va más allá de la mera recopilación de la información proveniente de diversas fuentes.
10.00	14.00	6.00



HORAS NO PRESENCIALES

AF2 - Ejercicios prácticos/resolución de problemas - Ejercicios y resolución de problemas planteados por el/a docente a partir de una breve lectura, un material preparado para la ocasión, o cualquier otro tipo de datos o informaciones que supongan un desafío intelectual para el/la estudiante.	AF3 - Trabajos individuales/grupales - Trabajos de carácter teórico o práctico, individuales o grupales, que implican la lectura de artículos, revistas, informes de investigación, capítulos de libros, informaciones en Internet, etc. y la redacción de una reflexión personal (de diverso calado y extensión) que va más allá de la mera recopilación de la información proveniente de diversas fuentes.	AF5 - Estudio personal y documentación - Estudio individual que el/la estudiante realiza para comprender, reelaborar y retener un contenido científico con vistas a una posible aplicación en el ámbito de su profesión. Lectura individual de textos de diferente tipo (libros, revistas, artículos sueltos, prensa, publicaciones en Internet, informes sobre experiencias prácticas, etc.) relacionados con las materias de estudio.
15.00	15.00	15.00
CRÉDITOS ECTS: 3,0 (75,00 horas)		

EVALUACIÓN Y CRITERIOS DE CALIFICACIÓN

El uso de IA para crear trabajos completos o partes relevantes, sin citar la fuente o la herramienta o sin estar permitido expresamente en la descripción del trabajo, será considerado plagio y regulado conforme al Reglamento General de la Universidad.

Actividades de evaluación	Criterios de evaluación	Peso
Comprensión de conceptos y resolución de problemas de tratamiento de datos.	Comprensión de conceptos y resolución de problemas de tratamiento de datos.	50 %
Diseño y ejecución de flujos de análisis de datos; calidad y documentación.	Diseño y ejecución de flujos de análisis de datos; calidad y documentación.	40 %
Asistencia al aula y participación razonada.	Asistencia al aula y participación razonada.	10 %

Calificaciones

Inasistencia y plagio:

La inasistencia comprobada e injustificada a más de un tercio de las horas lectivas impartidas en cada asignatura, contabilizadas en el período comprendido desde el primer día de clase hasta quince días antes del inicio del período de exámenes, puede tener como consecuencia la imposibilidad de presentarse a examen en la convocatoria ordinaria del mismo curso académico (artículo 93.1 del Reglamento General). En el supuesto de que se aplicara esta consecuencia, la pérdida de convocatoria se extenderá a la convocatoria extraordinaria.

El incurrir en una falta académica grave, como es el plagio de materiales previamente publicados, el copiar en su examen u otra actividad evaluada, o no respetar las normas de convivencia básicas puede llevar a la apertura de un expediente sancionador y la pérdida de dos convocatorias.



COMILLAS

UNIVERSIDAD PONTIFICIA

ICAI

ICADE

CIHS

GUÍA DOCENTE
2026 - 2027

Convocatoria ordinaria:

Los porcentajes que compondrán la nota final de la asignatura en convocatoria ordinaria se muestran en el anterior apartado de evaluación.

Convocatoria extraordinaria:

La calificación de la asignatura coincidirá íntegramente con la nota del examen oficial en dicha convocatoria.

Alumnos que han cumplido escolaridad y con dispensa de escolaridad:

El alumno debe informar a la profesora mediante correo electrónico al comienzo del curso (hasta un máximo de 20 días naturales desde el inicio de la asignatura) sobre su escolaridad cumplida o dispensa de escolaridad.

BIBLIOGRAFÍA Y RECURSOS

Bibliografía Básica

- Provost, F. y Fawcett, T. (2013). *Data science for business: What you need to know about data mining and data-analytic thinking*. O'Reilly Media.
- Torres, J., Macías, M., Gómez, M. y Tous, R. (2015). *Introducción a Apache Spark para programar el Big Data*. Editorial UOC.
- Feldman, R. y Sanger, J. (2007). *The text mining handbook: Advanced approaches in analyzing unstructured data*. Cambridge University Press.
- McKinney, W. (2022). *Python for data analysis* (3.ª ed.). O'Reilly Media.

Bibliografía Complementaria

- Bird, S., Klein, E. y Loper, E. (2009). *Natural language processing with Python*. O'Reilly Media.
- Vajjala, S., Majumder, B., Gupta, A. y Surana, H. (2020). *Practical natural language processing: A comprehensive guide to building real-world NLP systems*. O'Reilly Media.
- Rojo, G. (2021). *Introducción a la lingüística de corpus en español*. Routledge. <https://doi.org/10.4324/9781003119760>

En cumplimiento de la normativa vigente en materia de **protección de datos de carácter personal**, le informamos y recordamos que puede consultar los aspectos relativos a privacidad y protección de datos que ha aceptado en su matrícula entrando en esta web y pulsando "descargar"

<https://servicios.upcomillas.es/sedelectronica/inicio.aspx?csv=02E4557CAA66F4A81663AD10CED66792>