GUÍA DOCENTE

BÚSQUEDA Y ANÁLISIS DE LA INFORMACIÓN

DOBLE GRADO EN INGENIERÍA DEL SOFTWARE

MODALIDAD: PRESENCIAL

CURSO ACADÉMICO: 2025-2026

Denominación de la asignatura:	Búsqueda y Análisis de la Información
Titulación:	Doble Grado en Ingeniería del Software
Facultad o Centro:	Centro Universitario de Tecnología y Arte Digital
Materia:	Ingeniería de Datos
Curso:	3
Cuatrimestre:	2
Carácter:	OBM
Créditos ECTS:	6
Modalidad/es de enseñanza:	Presencial
Idioma:	Castellano
Profesor/a - email	Stanislav Vakaruk / stanislav.vakaruk@u-tad.com
Página Web:	http://www.u-tad.com/

DESCRIPCIÓN DE LA ASIGNATURA

Descripción de la materia

Los contenidos de la materia permiten a los alumnos comprender el flujo de búsqueda, ingesta, almacenamiento, procesamiento y análisis de información de datos y aproxima a los alumnos a las técnicas y tecnologías necesarias para la gestión de grandes cantidades de datos.

Descripción de la asignatura

Esta asignatura tiene como objetivo equipar a los estudiantes con técnicas avanzadas para extraer información de Internet, incluyendo el web scraping y la utilización de diversas APIs proporcionadas por servicios en línea.

Se enfatizan las aplicaciones prácticas en Ciencia de Datos utilizando entornos de código abierto, particularmente Python, que ofrece una amplia gama de bibliotecas para el análisis de datos.

Los componentes clave incluyen la preparación de datos (limpieza, consolidación y gestión básica de textos), minería de textos (minería de temas y análisis de sentimientos) y análisis de redes sociales (SNA) con un enfoque en la visualización.

Además, el curso cubre la aplicación de modelos de Deep Learning en el Procesamiento de Lenguaje Natural (NLP), específicamente los Modelos de Lenguaje Grande (LLMs) como GPT-3, utilizando marcos modernos como TensorFlow y PyTorch.

Al final del curso, los estudiantes tendrán experiencia práctica con técnicas modernas de extracción y análisis de datos, preparándolos para aplicaciones prácticas en varios campos de la Ciencia de Datos.

COMPETENCIAS Y RESULTADOS DE APRENDIZAJE DE LA MATERIA

Competencias (genéricas, específicas y transversales)

COMPETENCIAS BÁSICAS Y GENERALES

CB1: Que los estudiantes hayan demostrado poseer y comprender conocimientos en un área de estudio que parte de la base de la educación secundaria general, y se suele encontrar a un nivel que, si bien se apoya en libros de texto avanzados, incluye también algunos aspectos que implican conocimientos procedentes de la vanguardia de su campo de estudio.

CB2: Que los estudiantes sepan aplicar sus conocimientos a su trabajo o vocación de una forma profesional y posean las competencias que suelen demostrarse por medio de la elaboración y defensa de argumentos y la resolución de problemas dentro de su área de estudio.

CB3: Que los estudiantes tengan la capacidad de reunir e interpretar datos relevantes (normalmente dentro de su área de estudio) para emitir juicios que incluyan una reflexión sobre temas relevantes de índole social, científica o ética.

CB4: Que los estudiantes puedan transmitir información, ideas, problemas y soluciones a un público tanto especializado como no especializado.

CB5: Que los estudiantes hayan desarrollado aquellas habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía

CG1 - Capacidad para entender, planificar y resolver problemas a través del desarrollo de soluciones informáticas.

CG3 - Conocimiento de los fundamentos científicos aplicables a la resolución de problemas informáticos

CG4 - Capacidad para simplificar y optimizar los sistemas informáticos atendiendo a la comprensión de su complejidad

CG9 - Capacidad para aprender, modificar y producir nuevas tecnologías informáticas

CG10 - Uso de técnicas creativas para la realización de proyectos informáticos

CG11 - Capacidad de buscar, analizar y gestionar la información para poder extraer conocimiento de la misma

COMPETENCIAS ESPECÍFICAS

CE3 - Conocimiento del álgebra relacional y realización de consultas en lenguajes procedurales para el diseño de esquemas de

bases de datos normalizados basados en modelos de entidad-relación

CE10 - Capacidad para manejar un gestor de versiones de código y generar la documentación de una aplicación de forma

automática.

COMPETENCIAS TRANSVERSALES

CT1 - Conocimiento de la definición, el alcance y la puesta en práctica de los fundamentos de las metodologías de gestión de proyectos de desarrollo tecnológico

CT2 - Conocimiento de los principales agentes del sector y del ciclo de vida completo de un proyecto de desarrollo y comercialización de contenidos digitales

CT4 - Capacidad de actualización del conocimiento adquirido en el manejo de herramientas y tecnologías digitales en función del estado actual del sector y de las tecnologías empleadas

CT5 - Desarrollo de las habilidades necesarias para el emprendimiento digital.

Resultados de aprendizaje

Al acabar la titulación, el graduado o graduada será capaz de:

- Comprender e implementar los métodos de almacenamiento y administración eficaz en entornos distribuidos de datos no estructurados.

- Conocer y saber aplicar las distintas técnicas de aprendizaje supervisado, semi-supervisado y no supervisado.

- Entender y aplicar las técnicas de Deep learning

- Ser capaz de recuperar información mediante técnicas de web scraping o APIs normalizadas

- Entender y aplicar las técnicas de análisis del lenguaje natural

- Ser capaz de analizar contenidos de redes sociales

- Entender la naturaleza y representación de las imágenes digitales.

- Conocer las aplicaciones de las redes neuronales al análisis y generación de sonido, imagen estática y video.

- Desarrollar soluciones informáticas aplicadas a la visión por computador.

- Desarrollar un proyecto completo de datos aplicando metodología iterativa, desde el diseño hasta el despliegue.

CONTENIDO

Información textual. Modelos de relevancia y similaridad.

Búsqueda de información textual y no textual.

Búsqueda en la web.

Análisis de redes sociales.

TEMARIO

Tema 1. El ecosistema Python

Tema2. Minería de texto

Tema 3. Fuentes de datos de tipo texto

Tema 4. Modelización y análisis en minería de texto

Tema 5. Análisis de redes sociales

Tema 6. Aplicaciones de Deep Learning en NLP: LLMs

ACTIVIDADES FORMATIVAS Y METODOLOGÍAS DOCENTES

Actividades formativas

Actividad Formativa	Horas totales	Horas presenciales
Clases teóricas / Expositivas	29,375	29,375
Clases Prácticas	23,25	23,25
Tutorías	4	2
Estudio independiente y trabajo autónomo del alumno	50	0
Elaboración de trabajos (en grupo o individuales)	31,875	0
Actividades de Evaluación	5,25	5,25
Preparación y defensa del TFG	<<7- Preparación y defensa del TFG>>	<<Horas presenciales 7- Preparación y defensa del TFG>>

Metodologías docentes

Método expositivo o lección magistral

Aprendizaje de casos

Aprendizaje basado en la resolución de problemas

Aprendizaje basado en proyectos

Aprendizaje cooperativo o colaborativo

Aprendizaje por indagación

Metodología Flipped classroom o aula invertida

Gamificación

Just in time Teaching (JITT) o aula a tiempo

Método expositivo o lección magistral

Método del caso

Aprendizaje basado en la resolución de problemas

Aprendizaje basado en proyectos

Aprendizaje cooperativo o colaborativo

Aprendizaje por indagación

Metodología flipped classroom o aula invertida

Gamificación

DESARROLLO TEMPORAL

UNIDADES DIDÁCTICAS / TEMAS PERÍODO TEMPORAL

1. El ecosistema Python (1 semana)
2. Minería de texto (2 semanas)
3. Fuentes de datos de tipo texto (3 semanas)
4. Modelización y análisis en minería de texto (4 semanas)
5. Análisis de redes sociales (2 semanas)
6. Aplicaciones de Deep Learning en NLP: LLMs (2 semanas)

SISTEMA DE EVALUACIÓN

ACTIVIDAD DE EVALUACIÓN	VALORACIÓN MÍNIMA RESPECTO A LA CALIFICACIÓN FINAL (%)	VALORACIÓN MÁXIMA RESPECTO A LA CALIFICACIÓN FINAL (%)
Evaluación de la participación en clase, en prácticas o en proyectos de la asignatura	10	30
Evaluación de trabajos, proyectos, informes, memorias	40	80
Prueba Objetiva	10	60
Evaluación del TFG	<<4-(MIN)Evaluación del TFG>>	0

CRITERIOS ESPECÍFICOS DE EVALUACIÓN

ACTIVIDAD DE EVALUACIÓN	CONVOCATORIA ORDINARIA	CONVOCATORIA EXTRAORDINARIA
Evaluación de la participación en clase, en prácticas o en proyectos de la asignatura	10	10
Evaluación de trabajos, proyectos, informes, memorias	70	70
Prueba Objetiva	20	20
Evaluación del TFG	<<4-(MIN)Evaluación del TFG>>	0

Consideraciones generales acerca de la evaluación

Para aprobar la asignatura, los estudiantes deben cumplir con los siguientes criterios:

- Obtener una puntuación media mínima de 5 en las prácticas.
- Obtener al menos un 5 en el examen final.

Además, si un estudiante ha asistido a más del 80% de las clases y ha entregado al menos el 80% de los ejercicios, sus calificaciones pueden promediarse con una puntuación mínima de 4 en el examen final. Sin embargo, esto no implica que una puntuación de 4 en el examen final sea suficiente para aprobar.

Notas importantes:

- Cualquier instancia de copia o plagio resultará en un suspenso automático con una calificación de 0 en la tarea o examen en cuestión.

- El instructor se reserva el derecho de hacer preguntas sobre la tarea para verificar el trabajo individual de cada estudiante.

Fórmula para el Cálculo de la Calificación (Convocatoria Ordinaria)

La calificación final (CF) se calcula utilizando la siguiente fórmula:

$CF=(0.10\times E)+(0.70\times P)+(0.20\times EF)$

Donde:

- (E) = Participación en ejercicios
- (P) = Prácticas
- (EF) = Examen final

Por ejemplo, si un estudiante obtiene 8 en participación, 6 en prácticas y 5 en el examen final, la calificación final se calcularía de la siguiente manera:

$CF=(0.10\times 8)+(0.70\times 6)+(0.20\times 5)=0.8+4.2+1=6$

Fórmula para el Cálculo de la Calificación (Convocatoria Extraordinaria)

En el caso de una convocatoria extraordinaria, la calificación final se calcula utilizando la siguiente fórmula:

$CF=(0.70\times P)+(0.30\times EF)$

Donde:

- (P) = Prácticas
- (EF) = Examen final

Por ejemplo, si un estudiante obtiene 7 en prácticas y 6 en el examen final, la calificación final se calcularía de la siguiente manera:

$CF=(0.70\times 7)+(0.30\times 6)=4.9+1.8=6.$

BIBLIOGRAFÍA / WEBGRAFÍA

- Cheat-sheet Python: https://www.pythoncheatsheet.org/ ; https://www.datacamp.com/cheat-sheet/category/python
- Documentación oficial de Python: https://docs.python.org/es/3/tutorial/index.html
- Guía oficial de gestión de paquetes: https://packaging.python.org/en/latest/tutorials/installing-packages/
- Trabajar con CSV mediante módulo nativo de Python: https://docs.python.org/3/library/csv.html
- Trabajar con JSON mediante módulo nativo de Python: https://docs.python.org/3/library/json.html
- Parquet VS Avro: https://airbyte.com/data-engineering-resources/parquet-vs-avro

- Documentación de Polars: https://docs.pola.rs/
- Documentación de Pandas: https://pandas.pydata.org/docs/index.html
- Polars cheat-sheet: https://franzdiebold.github.io/polars-cheat-sheet/Polars_cheat_sheet.pdf
- Pandas cheat-sheet: https://www.datacamp.com/cheat-sheet/pandas-cheat-sheet-for-data-science-in-python
- Documentación oficial de re en Python: https://docs.python.org/3/library/re.html
- Tutorial de expresiones regulares (en inglés): https://www.regular-expressions.info/
- Herramientas online para testear expresiones: https://regex101.com/ ; https://regexr.com/
- Cheatsheet de expresiones regulares: https://www.datacamp.com/cheat-sheet/regular-expresso
- Repositorio y documentación oficial de la librería regex: https://pypi.org/project/regex/
- Listado de módulos de NLTK: https://www.nltk.org/py-modindex.html
- Instalación de datos de NLTK: https://www.nltk.org/data.html
- spaCy 101: https://spacy.io/usage/spacy-101
- spaCy universe: https://spacy.io/universe
- Comparativa de rendimientos: https://pola.rs/posts/benchmarks/
- BurpSuite community edition: https://portswigger.net/burp/communitydownload
- BeautifulSoup documentation: https://beautiful-soup-4.readthedocs.io/
- Selenium documentation: https://www.selenium.dev/documentation/
- Ética en scraping: https://www.meritdata-tech.com/resources/blog/data/web-scraping-best-practices-ethical-data-collection/
- Documentación oficial de gensim: https://radimrehurek.com/gensim/
- Modelado y visualización de tópicos: https://neptune.ai/blog/pyldavis-topic-modelling-exploration-tool-that-every-nlp-data-scientist-should-know
- SNA: https://github.com/ladamalina/coursera-sna/blob/master/Syllabus.pdf
- NetworkX para redes: https://networkx.org/documentation/stable/
- TensorFlow: https://www.tensorflow.org/guide?hl=es
- Pytorch: https://pytorch.org/tutorials/
- Keras: https://keras.io/guides/

MATERIALES, SOFTWARE Y HERRAMIENTAS NECESARIAS

Tipología del aula

Aula teórica

Equipo de proyección y pizarra

Materiales:

Ordenador personal

Software:

R última versión (https://cran.r-project.org/), en el momento de escribir esta guía,v. 4.3.2 (en general será necesaria versión >= 4.3.x)

y RStudio Desktop:(https://posit.co/download/rstudio-desktop/)

+ librerías stringr, tm, quanteda, tidyverse, tidytext, topicmodels, igraph (veremos suinstalación en las primeras clases)

Gephi (https://gephi.org/users/download/)

Entorno virtual python + keras + pytorch