# Guía docente
# 340459 - PEDT - Procesamiento y Explotación de Datos Textuales

**Última modificación:** 17/05/2023

**Unidad responsable:** Escuela Politécnica Superior de Ingeniería de Vilanova i la Geltrú
**Unidad que imparte:** 723 - CS - Departamento de Ciencias de la Computación.

**Titulación:** GRADO EN INGENIERÍA INFORMÁTICA (Plan 2018). (Asignatura optativa).

**Curso:** 2023 **Créditos ECTS:** 6.0 **Idiomas:** Catalán

## PROFESORADO

**Profesorado responsable:** Neus Català Roig

**Otros:**

## CAPACIDADES PREVIAS

- Saber aplicar los conceptos básicos de álgebra lineal, matemática discreta, probabilidad y estadística.
- Tener conocimientos básicos de programación con Python.
- Tener conocimientos básicos de los conceptos estudiados en las asignaturas de MIDA y REIN.

## REQUISITOS

Haber aprobado ESTA, AMEP y DABD o al menos estar matriculado/a.
Se recomienda haber cursado MIDA y REIN.

## COMPETENCIAS DE LA TITULACIÓN A LAS QUE CONTRIBUYE LA ASIGNATURA

**Específicas:**
I_CECO7. CECO7. Capacidad para conocer y desarrollar técnicas de aprendizaje computacional y diseñar e implementar aplicaciones y sistemas que las utilicen, incluyendo las dedicadas a extracción automática de información y conocimiento a partir de grandes volúmenes de datos.
I_CECO1. CECO1. Capacidad para tener un conocimiento profundo de los principios fundamentales y modelos de la computación y saberlos aplicar para interpretar, seleccionar, valorar, modelar, y crear nuevos conceptos, teorías, usos y desarrollos tecnológicos relacionados con la informática.
I_CECO4. CECO4. Capacidad para conocer los fundamentos, paradigmas y técnicas propias de los sistemas inteligentes y analizar, diseñar y construir sistemas, servicios y aplicaciones informáticas que utilicen dichas técnicas en cualquier ámbito de aplicación.
I_CEIS6. CEIS6. Capacidad para diseñar soluciones apropiadas en uno o más dominios de aplicación utilizando métodos de la ingeniería del software que integren aspectos éticos, sociales, legales y económicos.
I_CEIS4. CEIS4. Capacidad de identificar y analizar problemas y diseñar, desarrollar, implementar, verificar y documentar soluciones software sobre la base de un conocimiento adecuado de las teorías, modelos y técnicas actuales.

## METODOLOGÍAS DOCENTES

## OBJETIVOS DE APRENDIZAJE DE LA ASIGNATURA

Los datos textuales se encuentran en todas partes por ejemplo en libros, artículos, leyes, análisis financieras, registros médicos, redes sociales, etc. Se estima que representan entre el 80% y el 90% de los datos almacenados. Para poder extraer, resumir y analizar información a partir de grandes volúmenes de datos textuales se requieren métodos específicos. El campo conocido como Text Mining usa técnicas de computación para extraer información de datos textuales de forma automática.

La asignatura cubre los componentes básicos del Procesamiento del Lenguaje Natural (Natural Language Processing) y cómo se usan en las tareas de Minería de Textos (Text Mining). También se estudian algunas de sus aplicaciones como la Clasificación de Documentos, el Análisis de Sentimiento (Sentiment Analysis o Opinion Mining) y la Extracción de Información.

## HORAS TOTALES DE DEDICACIÓN DEL ESTUDIANTADO

| Tipo | Horas | Porcentaje |
|---|---|---|
| Horas grupo grande | 30,0 | 50.00 |
| Horas grupo pequeño | 30,0 | 50.00 |

**Dedicación total:** 60 h

## CONTENIDOS

### 1. Procesos de obtención de datos textuales

**Descripción:**
Obtener o construir un corpus. Creación de un corpus con datos extraídos de distintas fuentes: correos electrónicos, artículos de la Wikipedia, informes financieros, obras literarias o páginas web de interés. Rastreo de los datos (scrapping o web crawling).

**Actividades vinculadas:**
LABORATORIO
CUESTIONARIOS
PROYECTO

**Dedicación:** 6h 30m
Grupo grande/Teoría: 4h
Grupo pequeño/Laboratorio: 2h 30m

### 2. Preprocesamiento de datos textuales

**Descripción:**
Procesamiento sintáctico simple: limpieza de texto, normalización y tokenización.
Procesamiento lingüístico avanzado: desambiguación y etiquetado gramatical (part-of-speech tagging).

**Actividades vinculadas:**
LABORATORIO
CUESTIONARIOS
PROYECTO

**Dedicación:** 7h 30m
Grupo grande/Teoría: 4h
Grupo pequeño/Laboratorio: 3h 30m

### 3. Procesamiento del Lenguaje Natural: tareas principales y aplicaciones

**Descripción:**
Introducción al PLN. Tareas principales: etiquetado gramatical (PoS tagging), análisis sintáctico e interpretación semántica.
Algunas aplicaciones (demos): clasificación de documentos, clustering de documentos, análisis de sentimento, extracción de información, resumen automático, traducción automática.

**Actividades vinculadas:**
LABORATORIO
CUESTIONARIOS
PROYECTO

**Dedicación:** 13h
Grupo grande/Teoría: 6h
Grupo pequeño/Laboratorio: 7h

### 4. Herramientas de PLN y colecciones de datos para la minería de textos

**Descripción:**
Herramientas de PLN en Python: Scikit-Learn, Natural Language Toolkit (NLTK), Gensim, spaCy, NetworkX.
Colecciones de datos para la minería de textos accesibles on-line.

**Actividades vinculadas:**
LABORATORIO
CUESTIONARIOS
PROYECTO

**Dedicación:** 12h
Grupo grande/Teoría: 6h
Grupo pequeño/Laboratorio: 6h

### 5. Introducción a las redes neuronales y el aprendizaje profundo (Deep Learning) aplicado a la minería de textos

**Descripción:**
Vectorización de los textos: bag-of-words, tf-idf, word embeddings.
Redes neuronales. Aplicaciones del aprendizaje profundo (Deep Learning) a la minería de textos.

**Actividades vinculadas:**
LABORATORIO
CUESTIONARIOS
PROYECTO

**Dedicación:** 12h
Grupo grande/Teoría: 6h
Grupo pequeño/Laboratorio: 6h

**6. Presentaciones de los proyectos**

**Descripción:**
Presentaciones de los proyectos elaborados por los estudiantes.

**Actividades vinculadas:**
PROYECTO

**Dedicación:** 9h
Grupo grande/Teoría: 4h
Grupo pequeño/Laboratorio: 5h

## SISTEMA DE CALIFICACIÓN

- Evaluación de las actividades realizadas en las sesiones de laboratorio: 60%
- Realización y presentación pública de un trabajo de análisis sobre uno de los temas estudiados en el curso: 30%
- Cuestionarios: 10%

Dado que el 100% de la asignatura se evalúa a través de trabajos prácticos, no hay ningún control final de carácter global ni tampoco ningún control de reavaluación en forma de examen escrito.

## NORMAS PARA LA REALIZACIÓN DE LAS PRUEBAS.

Los informes de las actividades de laboratorio se entregan de forma no presencial en el plazo indicado para cada sesión.
Los cuestionarios son presenciales e individuales.

## BIBLIOGRAFÍA

**Básica:**
- Ignatow, Gabe; Mihalcea, Rada F. An Introduction to text mining : research design, data collection, and analysis. Thousand Oaks, California: SAGE Publications, Inc, 2018. ISBN 9781506337005.
- Jurafsky, Dan; Martin, James H. Speech and language processing : an introduction to natural language processing, computational linguistics, and speech recognition [en línea]. 3rd ed. Upper Saddle River: Els autors, 2019 [Consulta: 28/04/2022]. Disponible a: https://web.stanford.edu/~jurafsky/slp3/.