FICHA IDENTIFICATIVA

Datos de la Asignatura

Código	36429
Nombre	Procesado del Lenguaje Natural
Ciclo	Grado
Créditos ECTS	6.0
Curso académico	2020 - 2021

Titulación(es)

Titulación	Centro	Curso	Periodo
1400 - Grado de Ingeniería Informática	Escuela Técnica Superior de Ingeniería	4	Segundo cuatrimestre
1406 - Grado en Ciencia de Datos	Escuela Técnica Superior de Ingeniería	3	Segundo cuatrimestre

Materias

Titulación	Materia	Caracter
1400 - Grado de Ingeniería Informática	16 - Materia Optativa	Optativa
1406 - Grado en Ciencia de Datos	9 - Aprendizaje automático y minería de datos	Obligatoria

Coordinación

Nombre	Departamento
VILA FRANCES, JOAN	242 - Ingeniería Electrónica

RESUMEN

Actualmente, gran parte de los datos disponibles para el análisis están formados por información no estructurada en forma de textos en lenguaje natural. Entre esta información encontramos páginas web (Wikipedia, periódicos digitales, blogs) o redes sociales (Facebook, Twitter). Poder analizar estos textos, mediante algoritmos de procesado de lenguaje natural (PLN), resulta muy útil para que las organizaciones puedan tomar mejores decisiones.

Los algoritmos de aprendizaje automático no son capaces de entender texto o caracteres, por lo que el PLN realiza todo el procesado necesario para convertir estos datos en forma de texto en un formato entendible por las máquinas (números) y así poder realizar todo tipo de análisis posterior. Entre las aplicaciones más comunes del PLN se encuentran la clasificación de textos, búsqueda y extracción de información, traducción automática o sistemas de respuesta automática, entre otros.

Todos los pasos del PLN, desde la captura del texto en cualquier formato a la manipulación y análisis de este para obtener la información relevante, son abordados en la asignatura obligatoria 36429, Procesado de Lenguaje Natural que se imparte en el segundo cuatrimestre del tercer curso.

Las clases de teoría se impartirán en castellano y las clases prácticas y de laboratorio según consta en la ficha de la asignatura disponible en la web del grado.

CONOCIMIENTOS PREVIOS

Relación con otras asignaturas de la misma titulación

No se han especificado restricciones de matrícula con otras asignaturas del plan de estudios.

Otros tipos de requisitos

Se recomienda haber superado las asignaturas Modelos Lineales (segundo curso) y Aprendizaje Máquina (primer cuatrimestre de tercer curso).

COMPETENCIAS

1400 - Grado de Ingeniería Informática

- C1 - 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.
- C2 - Capacidad para adquirir, obtener, formalizar y representar el conocimiento humano en una forma computable para la resolución de problemas mediante un sistema informático en cualquier ámbito de aplicación, particularmente los relacionados con aspectos de computación, percepción y actuación en ambientes o entornos inteligentes.
- C3 - 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.

1406 - Grado en Ciencia de Datos

- Que los estudiantes hayan desarrollado aquellas habilidades de aprendizaje necesarias para emprender estudios posteriores con un alto grado de autonomía.
- (CG06) Capacidad de acceso y gestión de la información en diferentes formatos para su posterior análisis con el fin de obtener conocimiento a partir de datos.
- (CT04) Ser responsables de su propio desarrollo profesional y de su especialización, aplicando los conocimientos adquiridos en la identificación de salidas profesionales y yacimientos de empleo.

- (CT05) Capacidad para evaluar las ventajas e inconvenientes de diferentes alternativas metodológicas y/o tecnológicas en distintos ámbitos de aplicación.
- (CE03) Capacidad para resolver problemas de clasificación, modelización, segmentación y predicción a partir de un conjunto de datos.
- (CE07) Capacidad para modelar la dependencia entre una variable respuesta y varias variables explicativas, en conjuntos de datos complejos, mediante técnicas de aprendizaje máquina, interpretando los resultados obtenidos.

RESULTADOS DE APRENDIZAJE

Saber segmentar texto en elementos simples (CG06)

Conocer las técnicas de procesado de lenguaje natural (CG06, CT04, CE03, CE07).

Conocer e implementar las aplicaciones más extendidas de procesado de lenguaje natural (CB05, CT05, CE03, CE07).

Como consecuencia de los resultados de aprendizaje adquiridos, el alumnado adquirirá las siguientes destrezas:

- Ser capaz de cargar y tratar texto en Python.
- Saber usar expresiones regulares sobre texto.
- Conocer y saber usar las librerías más importantes de PLN para Python.
- Ser capaz de convertir texto en vectores numéricos para su posterior tratamiento con algoritmos de aprendizaje máquina.
- Ser capaz de diseñar una aplicación de clasificación de texto.
- Ser capaz de analizar grandes volúmenes de texto para extraer sus temáticas más representativas y realizar búsquedas de información.

DESCRIPCIÓN DE CONTENIDOS

1. Introducción al Procesado de Lenguaje Natural

1.1. Qué es el PLN
1.2. La importancia del texto
1.3. Aproximaciones históricas al PLN
1.4. Flujo de trabajo en aplicaciones de PLN

2. Uso de texto en Python

2.1. Cadenas de texto en Python
2.2. Expresiones regulares
2.3. Carga de texto
2.4. Librerías de PLN en Python
2.5. Captura de contenidos web (Web scraping)

3. Preprocesado de texto

3.1. Limpieza y normalización del texto
3.2. División de texto (tokens)
3.3. Análisis morfológico (lemmas)
3.4. Análisis semántico (Part of Speech)
3.5. Análisis gramatical (dependencias)

4. Extracción de características

4.1. Modelos de espacio de vectores
4.2. Modelo Bag of Words
4.3. Modelo TF-IDF
4.4. Vectores de palabra (Word embeddings)
4.5. Vectores de documento

5. Aplicaciones del NLP

5.1. Clasificación
5.2. Extracción de información
5.3. Búsqueda de entidades
5.4. Modelado de temática
5.5. Modelos secuenciales

6. Prácticas de Procesado de Lenguaje Natural

En este bloque se realizarán una serie de ejercicios prácticos para aplicar los conceptos de PLN en el aula informática.
Práctica 0. Uso de texto en Python
Práctica 1. Expresiones regulares
Práctica 2. Web scraping
Práctica 3. Uso de librerías PLN en Python
Práctica 4. Preprocesado de texto en Python
Práctica 5. Extracción de características
Práctica 6. Modelado de temática
Práctica 7. Aplicación práctica: análisis de sentimientos

VOLUMEN DE TRABAJO

ACTIVIDAD	Horas	% Presencial
Clases de teoría	30,00	100
Prácticas en laboratorio	20,00	100
Prácticas en aula	10,00	100
Elaboración de trabajos en grupo	10,00	0
Elaboración de trabajos individuales	10,00	0
Estudio y trabajo autónomo	15,00	0
Lecturas de material complementario	5,00	0
Preparación de actividades de evaluación	10,00	0
Preparación de clases de teoría	10,00	0
Preparación de clases prácticas y de problemas	10,00	0
Resolución de casos prácticos	15,00	0
Resolución de cuestionarios on-line	5,00	0
TOTAL	**150,00**	

METODOLOGÍA DOCENTE

Las clases combinarán el contenido teórico con el práctico

MD1 - Actividades teóricas. Desarrollo expositivo de la materia con la participación del alumnado en la resolución de cuestiones puntuales. Realización de cuestionarios individuales de evaluación.

En las actividades teóricas de carácter presencial se desarrollarán los temas de la asignatura proporcionando una visión global e integradora, analizando con mayor detalle los aspectos clave y de mayor complejidad, fomentando, en todo momento, la participación del alumnado (CB05, CT05).

MD2 - Actividades prácticas. Aprendizaje mediante resolución de problemas, ejercicios y casos de estudio a través de los cuales se adquieren competencias sobre los diferentes aspectos teóricos de la materia. (CB05, CG06, CE03, CE07)

Las actividades teóricas se complementan con prácticas de laboratorio con el objetivo de poner en uso los conceptos básicos y ampliarlos con el conocimiento y la experiencia que se vayan adquiriendo durante la realización de los trabajos propuestos.

MD4 -Trabajos en laboratorio y/o aula ordenador. Aprendizaje mediante la realización de actividades guiadas desarrolladas de forma individual o en grupos reducidos y llevadas a cabo en laboratorios y/o aulas de ordenador. (CB05, CG06, CT04, CE03, CE07)

Además de las actividades presenciales, los estudiantes deberán realizar tareas personales (fuera del aula) sobre: cuestiones y problemas, así como la preparación de clases y exámenes (estudio). Estas tareas se realizarán principalmente de manera individual, con el fin de potenciar el trabajo autónomo, pero adicionalmente se incluirán trabajos, especialmente la preparación y resolución de prácticas de laboratorio, que requieran la participación de pequeños grupos de estudiantes (2-3) para fomentar la capacidad de integración en grupos de trabajo.

Se utilizará la plataforma de e-learning (Aula Virtual) de la Universitat de València como soporte de comunicación con el alumnado. A través de ella se tendrá acceso al material didáctico utilizado en clase y los guiones de las prácticas de laboratorio, así como los problemas y ejercicios a resolver.

EVALUACIÓN

La evaluación del aprendizaje de los conocimientos y competencias conseguidas por los estudiantes se hará de forma continuada a lo largo del curso, y constará de los siguientes bloques:

- SE1 - Prueba objetiva, consistente en un examen que consta tanto de cuestiones teórico-prácticas como de problemas (evaluación de competencias CB05, CT05, CE03, CE07) (50%) (Nota: Todos los porcentajes están referidos a la nota final)
    - SE1-1 (40%) Examen de teoría-problemas
    - SE1-2 (10%) Examen de laboratorio
- SE2 - Evaluación de las prácticas de laboratorio a partir de la elaboración de trabajos/memorias y/o exposiciones orales (evaluación de competencias CB05, CG06, CT04, CE03, CE07) (30%)
    - SE2-1 (20%) Realización de un mini proyecto consistente en el desarrollo de una aplicación completa de PLN para la clasificación de textos
    - SE2-2 (10%) Asistencia y evaluación de las sesiones de laboratorio (Actividad NO RECUPERABLE)
- SE3 - Evaluación continua de cada alumno. (20%)
    - SE3-1 (5%) Asistencia regular a las actividades presenciales previstas (evaluación de competencias CB05). (Actividad NO RECUPERABLE)
    - SE3-2 (15%) Resolución de cuestiones y problemas propuestos (evaluación de competencias CB05, CG06, CE03). (Actividad NO RECUPERABLE)

La nota final de la asignatura se calculará como la media ponderada de cada uno de los apartados anteriores, de acuerdo con el siguiente criterio: SE-1 (50%), SE-2 (30%), SE-3 (20%).

Consideraciones particulares sobre la evaluación:

• Es necesario obtener una calificación mínima de 4 (sobre 10) en cada uno de los apartados de evaluación SE1-1, SE1-2 y SE2-1.

• Las actividades SE2-2, SE3-1 y SE3-2 no son recuperables.

En cualquier caso, el sistema de evaluación se regirá por lo establecido en el Reglamento de Evaluación y Calificación de la Universidad de Valencia para Grados y Másteres:

https://webges.uv.es/uvTaeWeb/MuestraInformacionEdictoPublicoFrontAction.do?accion=inicio&idEdicto Seleccionado=5639

REFERENCIAS

Básicas

- Sohom Ghosh, Dwight Gunning. Natural Language Processing Fundamentals. Packt Publishing, 2019.
- Akshay Kulkarni, Adarsha Shivananda. Natural Language Processing Recipes: Unlocking Text Data with Machine Learning and Deep Learning using Python. Apress, 2019 (disponible e-libro)
- Dipanjan Sarkar. Text Analytics with Python: A Practitioner's Guide to Natural Language Processing. Apress 2019 (disponible e-libro)
- Steven Bird, Ewan Klein, Edward Loper. Natural Language Processing with Python. OReally Media, 2009.

Complementarias

- Jacob Eisentein. Natural Language Processing. 2018 (disponible bajo licencia CC-BY-NC-ND)
- Sowmya Vajjala, Bodhisattwa Majumder, Anuj Gupta, Harshit Surana. Practical Natural Language Processing. OReally Media, 2020

ADENDA COVID-19

Esta adenda solo se activará si la situación sanitaria lo requiere y previo acuerdo del Consejo de Gobierno

La metodología docente de la asignatura seguirá el Modelo Docente aprobado por la Comisión Académica de Título de Ciencia de Datos (https://go.uv.es/cienciadatos/ModelDocentGCD2Q). En caso de que se produzca un cierre de las instalaciones por causas sanitarias que afecte total o parcialmente a las clases de la asignatura, estas serán sustituidas por sesiones no presenciales siguiendo los horarios establecidos. Si el cierre afectara a alguna prueba de evaluación presencial de la asignatura, esta será sustituida por una prueba de naturaleza similar que se realizará en modalidad virtual a través de las herramientas informáticas soportadas por la Universitat de València. Los porcentajes de cada prueba de evaluación permanecerán invariables, según lo establecido por esta guía.