A10. Big Data: Herramientas para el procesamiento de datos masivos

MÁSTER UNIVERSITARIO EN INVESTIGACIÓN EN INTELIGENCIA ARTIFICIAL

UNIVERSIDAD INTERNACIONAL MENÉNDEZ PELAYO

DATOS GENERALES

Breve descripción

En asignatura se aprenderán las principales herramientas disponibles para trabajar con bases de datos masivas. Veremos las diferentes técnicas y algoritmos de analítica de datos, desde el preprocesado a la clasificación, tratamiento de datos en streaming, etc.

Título asignatura

A10. Big Data: Herramientas para el procesamiento de datos masivos

Código asignatura

102473

Curso académico

2022-23

Planes donde se imparte

MÁSTER UNIVERSITARIO EN INVESTIGACIÓN EN INTELIGENCIA ARTIFICIAL

Créditos ECTS

9

Carácter de la asignatura

OPTATIVA

Duración

Anual

Idioma

Castellano

CONTENIDOS

Contenidos

Los avances tecnológicos de los últimos años han modificado nuestros hábitos y estilos de vida de una manera difícil de imaginar hace pocos años. El desarrollo de internet y su popularidad a nivel mundial han ayudado a eliminar fronteras y han creado multitud de servicios donde los datos transmitidos son un eje central de su funcionamiento. No obstante, estos datos no sólo se encuentran ligados a internet o a las redes sociales, sino que son parte fundamental de numerosas aplicaciones, tales como las colecciones de datos que nos proporcionan los instrumentos científicos, las redes de sensores, los dispositivos móviles, las transacciones comerciales, la genómica y la biomedicina, o los sistemas de información de la empresa.

Esta gran cantidad de datos disponible en la actualidad y las tecnologías necesarias para su procesamiento conforma lo que conocemos hoy día como "big data". Esta materia se centrará en el procesamiento de datos masivos, tanto en los principios formales como en las herramientas específicas para tratar estos volúmenes de datos.

- Big data.
- Procesamiento de datos masivos.
- Deep learning.
- Herramientas para el tratamiento de grandes volúmenes de datos: Hadoop, Spark, Mahout, MLLib.

Unidades

1. Módulo 1: Fundamentos de Big Data. Algunas aplicaciones
2. Módulo 2: Modelo de programación MapReduce
3. Módulo 3: Hadoop. Un caso de estudio
4. Módulo 4: Analítica para Big data. Generalidades y herramientas
5. Módulo 5: Algoritmos de Preprocesamiento
6. Módulo 6: Algoritmos de clasificación
7. Módulo 7: Algoritmos de Asociación
8. Módulo 8: Data streaming

9. Módulo 8: Herramientas Big data

COMPETENCIAS

Generales

CG1 - Entender los conceptos, los métodos y las aplicaciones de la inteligencia artificial.

CG2 - Evaluar nuevas herramientas computacionales y de gestión del conocimiento en el ámbito de la Inteligencia Artificial.

CG3 - Gestionar de manera inteligente los datos, la información y su representación.

Específicas

CE2 - Aplicar las técnicas de aprendizaje automático utilizando la metodología de validación y presentación de resultados más apropiada en cada caso.

CE5 - Analizar las fuentes documentales propias del ámbito de la investigación en Inteligencia Artificial para poder determinar cuáles de ellas son relevantes en la resolución de problemas concretos.

PLAN DE APRENDIZAJE

Actividades formativas

A1 - **Sesiones presenciales virtuales:**

- **(clases en vídeo)**: visionado del material audiovisual que constituye parte de las lecciones de la asignatura. Se asume 1.5 veces el tiempo real del vídeo, ya que el estudiante deberá parar, repetir, etc. algunas partes del video (7 horas).

- **(guías de estudio)**: lectura del material escrito que constituye parte de las lecciones de la asignatura. Existen ejercicios simples con los que el estudiante puede practicar (19 horas).

A2 - **Trabajos individuales**: realización de los cuestionarios de evaluación (10 horas).

A3 - **Trabajo autónomo**: lectura del material relacionado con las prácticas, configuración del entorno de trabajo y realización de prácticas individuales y del ejercicio de investigación (145 horas).

A4 - **Foros y chats** y A5 **- Tutorías**: consultas y resolución de dudas, aclaraciones, ampliaciones y, en general, cualquier ayuda que se le pueda ofrecer al alumno durante la realización del curso (44 horas).

Puede consultar en este enlace el **Cronograma de Carga de Trabajo**.

SISTEMA DE EVALUACIÓN

Descripción del sistema de evaluación

E1 - **Valoración de los cuestionarios de evaluación**: los estudiantes realizarán seis breves cuestionarios de evaluación (test) para valorar sus conocimientos sobre ciencia de datos, Hadoop, Spark y sistemas de streaming. El peso en la nota final de este apartado será del 10 % del total. Además, deberán entregar un artículo de investigación individual en el que revisen un tema propuesto relacionado con los contenidos de la asignatura. El peso en la nota final de este apartado será del 20% del total (peso total de E1: 30%).

E2 - **Valoración de la participación en foros y chats**: se valorará el nivel de participación/debate de los estudiantes que contará para la nota final (10%).

E3 - **Valoración de los trabajos individuales**: los estudiantes realizarán tres proyectos individuales de programación para valorar su capacidad de utilizar Spark batch, Spark streaming y Flink. El peso en la nota final de este apartado será del 60% del total.

Calendario de exámenes

Para la **convocatoria ordinaria**, habrá 3 fechas de entrega de trabajos final de curso. Los alumnos podrán entregar sus trabajos en cualquier momento, pero sólo en estas fechas se recogerán y evaluarán los que se hayan entregado. Las fechas serán:

• 13/01/23
• 17/03/23
• 31/05/23

Habrá una **convocatoria extraordinaria** en todas las asignaturas. Para su evaluación, la fecha límite para la entrega de trabajos será:

• 14/07/23

Las actas de la convocatoria ordinaria se cerrarán en julio de 2023 y las de la convocatoria extraordinaria en septiembre de 2023.

PROFESORADO

Profesor responsable

Alonso Betanzos, María Amparo

Catedrática de Ciencias de la Computación e Inteligencia Artificial
Universidad de A Coruña

Profesorado

Martínez Rego, David

Doctor en Informatica
Universidad de A Coruña

Cancela Barizo, Brais

Investigador Postdoctoral
Universidad de A Coruña

Novoa Paradela, David

Investigador predoctoral en grupo LIDIA-CITIC
Universidad de A Coruña

Eiras Franco, Carlos

Profesor Ayudante Doctor
Universidad de A Coruña

Bolón Canedo, Verónica

Profesora Titular G.I. Laboratorio de Investigación y Desarrollo en Inteligencia Artificial
Universidade da Coruña
Academia Joven de España y Real Academia de Ciencias Exactas, Físicas y Naturales

Pérez Freire, Carlos Javier

Teaching Assistant
Ingeniero en Informática/DEA en Telemática

Area Manager

HORARIO

Horario

Todas las asignaturas estarán en la plataforma a disposición de los estudiantes desde octubre hasta julio.

BIBLIOGRAFÍA Y ENLACES RELACIONADOS

Bibliografía

Sean T. Allen, Matthew Jankowski, and Peter Pathirana. *Storm Applied*. Manning 2015

Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia. *Learning Spark*. O'Reilly 2015

Sameer B. Wadkar, Hari Rajaram. *Flink in Action*. Manning 2017

Paul Butcher. *Seven concurrency models in seven weeks*. The Pragmatic Programmer 2014

Mahmoud Parsian. *Data Algorithms: Recipes for Scaling Up with Hadoop and Spark*. O'Reilly 2015

Tom White. *Hadoop: The Definitive Guide*, 4th Edition. O'Reilly 2015

Thilina Gunarathne. *Hadoop MapReduce v2 Cookbook*, 2nd Edition. Packt Publishing, 2015

Holden Karau, Andy Konwinski, Patrick Wendell, Matei Zaharia. *Learning Spark Lightning-Fast Big Data Analysis*. O'Reilly Media, 2015

Venkat Ankam. *Big Data Analytics.* Packt Publishing, 2016

Vladimir Bacvanski. *Introduction to Big Data An Overview of Fundamental Big Data Concepts, Tools, Techniques and Practices*. O'Reilly Media, 2015

Verónica Bolón-Canedo, Noelia Sánchez-Maroño, and Amparo Alonso-Betanzos. *Feature selection for high-dimensional data*. Springer, 2015

Salvador García, Julián Luengo, and Francisco Herrera. *Data preprocessing in data mining*. New York: Springer, 2015

Isabelle Guyon, Steve Gunn, Masoud Nikravesh, and Lofti Zadeh, Eds. *Feature Extraction: Foundations and Applications*. Springer, 2006