Ciencia de Datos Avanzada

Datos no estructurados y GenAI multimodal

Descripción del curso

Ciencia de Datos II es la continuación directa de Ciencia de Datos I. El curso se centra en el análisis, procesamiento e integración de **datos no estructurados**, incluyendo texto libre, datos web, documentos PDF, imágenes y audio.

Asimismo, el curso introduce el uso aplicado y crítico de **modelos fundacionales y GenAI multimodal**, a través de una infraestructura controlada por el instructor que limita el consumo de recursos mediante un sistema de créditos. El énfasis del curso está en el diseño de flujos de análisis reproducibles, responsables y alineados con problemas reales de ciencia de datos.

Objetivo general

Que el estudiante sea capaz de diseñar e implementar pipelines de análisis de datos no estructurados, integrarlos con datos estructurados y utilizar herramientas modernas de GenAI de forma crítica, controlada y orientada a la extracción de información.

Temario por semana

Semana 1

Introducción a los datos no estructurados.

- Qué son los datos no estructurados.
- Tipos: texto, web, documentos, imágenes y audio.
- Diferencias con datos estructurados.
- Casos reales en ciencia de datos.

Semana 2

Texto libre y preparación para análisis.

- Limpieza y normalización de texto.
- Tokenización y stopwords.
- Representaciones clásicas (TF-IDF).
- Introducción conceptual a embeddings.

Semana 3

Obtención de datos desde la web.

- Ética y legalidad del web scraping.
- Estructura básica de HTML.
- Scraping estático con `requests` y `BeautifulSoup`.

Semana 4

Scraping avanzado y datos semi-estructurados.

- JSON y XML.
- Scraping dinámico (Selenium / Playwright).
- Normalización y almacenamiento en PostgreSQL.

Semana 5

Documentos PDF y tablas.

- PDFs digitales vs escaneados.
- Extracción de tablas con Tabula.
- Limpieza y validación de datos documentales.

Semana 6

Texto en documentos.

- Extracción de texto desde PDFs.
- Problemas comunes de OCR.
- Integración de texto documental con datos estructurados.

Semana 7

Fundamentos de imágenes digitales.

- Qué es una imagen digital.
- Resolución, canales y espacios de color.
- Introducción a OpenCV.

Semana 8

Procesamiento clásico de imágenes.

- Filtros y transformaciones.
- Detección de bordes.
- Segmentación y contornos.

Semana 9

Extracción y análisis de información visual.

- OCR y extracción de texto desde imágenes.
- Conversión de información visual en datos estructurados.
- Análisis semántico de imágenes con modelos multimodales.
- Generación de imágenes como datos sintéticos.

Semana 10

Introducción a modelos fundacionales y LLMs.

- Qué es un LLM.
- Capacidades y limitaciones.
- Generación controlada de texto.
- Miscelánea: generación de audio a partir de texto (TTS).

Semana 11

Análisis avanzado de texto con GenAI.

- Clasificación automática de texto.
- Resumen de documentos largos.
- Reconocimiento y extracción de entidades (NER).
- Structured outputs y generación de JSON.

Semana 12

Retrieval-Augmented Generation (RAG).

- Motivación y arquitectura de RAG.
- Uso de embeddings para recuperación de contexto.
- Integración con PostgreSQL y documentos.
- Comparación: respuestas con y sin RAG.

Semana 13

Audio como dato no estructurado.

- Introducción al análisis de audio.
- Transcripción automática (ASR / Whisper).
- Limpieza y análisis de texto transcrito.
- Extracción de entidades desde audio.

Semana 14

Agentes, herramientas y control.

- Tool calling y funciones.
- Agentes simples para análisis de datos.
- Moderación y filtrado de contenido.
- Uso eficiente de créditos y batch processing.

Semana 15

Proyecto final.

- Definición del problema.
- Selección de fuentes no estructuradas.
- Diseño del pipeline de análisis.

Semana 16

Proyecto final.

- Desarrollo e integración final.
- Presentación de resultados.
- Discusión de alcances, errores y limitaciones.

Nota final

El uso de modelos de GenAI estará mediado por una plataforma de control implementada por el instructor, que limita el número de interacciones permitidas y promueve un uso consciente, responsable y estratégico de los recursos disponibles.