

# MÓDULO 9: TEXT MINING Y WEB MINING

## CONTENIDO

### 1. Introducción. Definición de conceptos

### 2. Análisis lingüístico

- 2.1. Análisis léxico (tokenization), etiquetado gramatical (parts of speech tagging, POS tagging), lematización (lemmatization) y análisis sintáctico (dependency parsing).

### 3. Análisis de texto

- 3.1. Técnicas para extraer palabras clave: n-gramas.
- 3.2. Coocurrencias
- 3.3. Matriz Documento-Término (DTM). Correlaciones

### 4. Representaciones distribuidas (embeddings)

- 4.1. Word embeddings (representación vectorial de palabras). Modelos: Word2Vec, GloVe, fastText.
- 4.2. Doc embeddings (representación vectorial de documentos). Modelo Doc2Vec.

### 5. Clasificación de texto

- 5.1. Aprendizaje supervisado: Técnicas y algoritmos de machine learning para clasificar documentos.
- 5.2. Aprendizaje no supervisado: Clustering de documentos.

### 6. Grandes Modelos de Lenguaje (LLM, Large Language Model)

- 6.1. La arquitectura Transformers. Modelos “Hugging Face” (implementación con PyTorch).
- 6.2. Tipos de modelos: Modelos de autocodificación (encoder models) o modelos de codificador bidireccional, Modelos secuencia a secuencia (encoder-decoder models) y Modelos autorregresivos (decoder models) o modelos de lenguaje causal.
- 6.3. Modelos orientados a tareas de procesamiento del lenguaje natural (NLP): clasificación de texto, análisis de sentimiento, zero-shot, reconocimiento de nombres de entidades (named entity recognition, NER), búsqueda de respuestas (question answering, QA), modelado de lenguaje, búsqueda semántica, resumen de texto, traducción, generación de texto, ...
- 6.4. Inteligencia Artificial Generativa.  
Modelos de generación de texto, imágenes, audio, vídeo, visión artificial, ...  
Modelos gpt-oss, Llama, DeepSeek, Qwen, Gemma, Mistral, LLaVA, Gemini, Vertex AI, Claude, ...  
Trabajar con LLM en local.  
Trabajar con LLM alojados en la nube.

### 7. Web Scraping. Web Crawling

El “Módulo 9” es fundamentalmente práctico:

- Programación en R.
- Grandes Modelos de Lenguaje (LLM): Programación en Python y R.

## PROFESORES RESPONSABLES

- Dr. Víctor Diego Fresno Fernández
- Dr. Álvaro Rodrigo Yuste
- D. Juan José García Muñoz

(Para mayor información de los profesores, puede consultar el menú “Equipo Docente” de la Web)