Laboratorio de Procesamiento del Lenguaje Natural

Lab 1: Adquisición de Datos mediante Web Scraping

Flujo de Trabajo

Organización de Archivos: Por ejemplo ‘Laboratorio_PLN’. Dentro, una carpeta para esta primera práctica (ej: 001_Clase_WebScraping).

GitHub: qué es y cómo crear un repositorio personal para la materia. El objetivo de hoy es solo crearlo.

Editor de Código (VS Code / Cursor): Reconocimiento de la interfaz.

Gestores de Paquetes (PIP / UV): son como una "Store" para librerías de Python.

Técnicas de Procesamiento del Habla

Web Scraping con Beautiful Soup: Extracción de Datos de la Web

¿Por qué Web Scraping?

- En el corazón de todo proyecto de NLP y Deep Learning están los datos. Los modelos necesitan grandes cantidades de datos para aprender y realizar tareas complejas. Pero ¿de dónde obtenemos estos datos?

- En muchos casos, los datos que necesitamos no están disponibles en formatos estructurados (como archivos CSV o bases de datos). En cambio, están esparcidos por la web en forma de páginas HTML.

Web Scraping como herramienta para la extracción de datos:

- ***El web scraping es la técnica de extraer información de sitios web de forma automatizada.*** Nos permite convertir la vasta cantidad de datos no estructurados en la web en un formato que nuestros algoritmos de NLP y Deep Learning puedan comprender y utilizar.

Ejemplos de aplicaciones reales:

- Analizar el sentimiento de los comentarios de clientes sobre un producto. ¿Cómo obtendrían esos datos…? Web scraping.

- Crear un sistema para resumir noticias automáticamente ¿De dónde sacarían el texto de las noticias? Web scraping.

- Construir una base de datos con información de productos de la competencia para análisis de precios ¿Cómo lo harían? Web scraping.

Buenas Prácticas y Cumplimiento

- Respetar robots.txt y Términos de Servicio.
- Evitar extraer PII sin consentimiento legal.
- Rate limiting, backoff, caché; no sobrecargar sitios.
- Preferir APIs si existen.
- Documentar orígenes, timestamps y licencias.

Herramientas Clave

- HTTP: Requests/HTTPX
- Parsing: BeautifulSoup (bs4), selectores CSS
- Dinámico: Playwright/Selenium (renderizado y scraping avanzado)
- Extracción de “artículo”: trafilatura/readability

Patrón Básico de Scraping

- Identificar URL y estructura.
- Descargar HTML (con headers/User-Agent).
- Parsear y seleccionar elementos con selectors CSS.
- Limpiar texto.
- Serializar a JSON/CSV.
- Caché y logs de acceso.

Patrón para Sitios Dinámicos

- Esperar selectores clave tras carga JS.
- Paginación, scroll infinito, manejar sesión/cookies.
- Captura y manejo de errores.

```
<html>
<head><title>Titulo de pagina</title>
</head>
<body>
    <p> Soy un parrafo</p>
    <div>Soy un texto en un DIV</div>
    <table><tr><td>soy una celda dentro de una tabla</td></tr>
    </table>
</body>
</html>
```

Glosario

Términos Clave para Web Scraping

PII

(Personally Identifiable Information / Información Personal Identificable)

- Datos que permiten identificar o contactar a una persona específica.
- Ejemplos: nombre completo, dirección, DNI, teléfono, e-mail, fotos del rostro.
- En scraping: No se deben recolectar ni almacenar PII sin base legal o consentimiento.

robots.txt

- Archivo ubicado en la raíz de un sitio web (ej: www.ejemplo.com/robots.txt).
- Indica a los bots qué partes del sitio pueden ser rastreadas.
- Es una referencia ética y técnica; ignorarlo puede violar normas del sitio.

Backoff

(espera adaptativa)

- Estrategia para espaciar los intentos tras un fallo o rechazo (por ejemplo, un rate limit).
- Puede ser exponencial (aumenta el tiempo: 2s → 4s → 8s…).
- Evita sobrecargar el servidor y mejora la estabilidad del scraping.

Timestamps

(marcas de tiempo)

- Registro de la fecha y hora exacta en que se obtiene un dato o se ejecuta un proceso.
- Ej: 2025-08-12 15:30:00 (GMT-3).
- Facilitan trazabilidad, replicar procesos y controlar actualizaciones.

Laboratorio de
Técnicas de Procesamiento de Habla

Profesor Matías Barreto

Encuentro 1

Introducción al Laboratorio

Formato del Laboratorio

Martes 6:30 PM - 8:30 PM -> TEO

Jueves 6:30 PM - 10:30 PM -> LAB

Evaluaciones Presenciales en formato Laboratorio

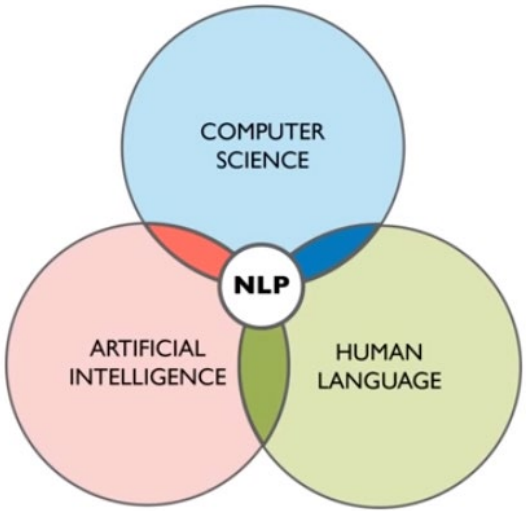
- Martes 25/9
- Jueves 20/11

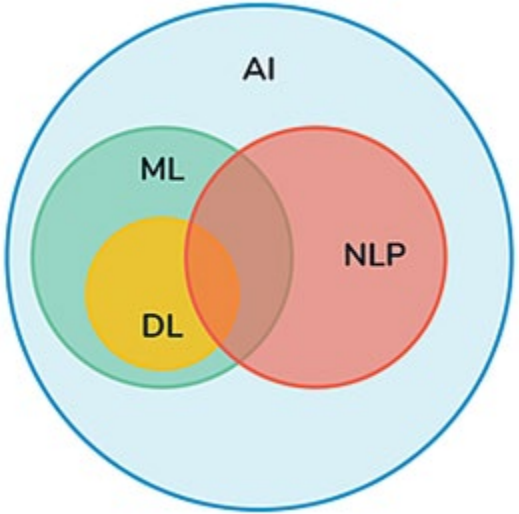
Clases y Materiales disponibles en Classroom

Criterios de Aprobación

- **Dominio del tema:** Se evalúa la comprensión teórica de los conceptos.
- **Aplicación práctica:** Se evalúa la capacidad de aplicar la teoría a la práctica.
- **Calidad del trabajo:** Se evalúa la presentación formal del trabajo, la cual incluye el desarrollo de prototipos funcionales de aplicaciones junto a documentación y, en algunos casos, elaboración de ensayos.
- **Trabajo en equipo:** Se evalúa la capacidad de colaborar y trabajar en equipo.
- **Presentación:** Se evalúa la claridad, la organización y el impacto comunicacional de la presentación.

Procesamiento del Lenguaje Natural (PLN)





PNL: es la parte de la Ciencias de la Computación e Inteligencia artificial que trabaja con lenguajes humanos

¿Qué es NLP?

- **Procesamiento de texto producido por humanos:** NLP se centra en algoritmos, tareas y problemas que toman texto generado por humanos como entrada.
- **Generación de información útil:** Produce información útil como etiquetas, representaciones semánticas, traducciones, resúmenes y texto generado.
- **Enfoque en la utilidad:** El objetivo es producir resultados útiles por sí mismos (ej. una traducción) o como entrada para otras tareas (ej. análisis sintáctico).
- **Abordaje científico:** NLP es un enfoque científico y basado en principios para lidiar con las complejidades y sutilezas de los lenguajes naturales.
- **Tratamiento de la ambigüedad:** Reconoce y aborda la ambigüedad inherente en los lenguajes naturales, tanto a nivel sintáctico como semántico.

Lenguaje Natural: El Desafío de la No Estructura

Puntos Clave:

- Datos no estructurados: No existe un modelo predefinido.
- Sin orden natural: No hay una forma inherente de organizar la información.
- Ambiguo e Intuitivo: El significado depende del contexto y la interpretación humana.
- Implicancia: la representación es clave

¿Qué NO es NLP?

- **No es solo una serie de "ifs" y "thens":** NLP no se trata de simplemente aplicar una serie de reglas condicionales básicas al texto.
- **No es una solución rápida y fácil:** No se trata de "ejecutar un compilador o un intérprete" en el texto y obtener una comprensión instantánea.
- **No es ignorar las complejidades del lenguaje:** NLP no ignora las complejidades y sutilezas de los lenguajes humanos; las aborda de manera sistemática.
- **No es adivinar o "intuir" el significado:** Requiere un enfoque más estructurado y fundamentado en principios lingüísticos y computacionales.
- **Tampoco es solo “usar un LLM”**: se requieren datos, evaluación y consideraciones éticas/legales

El **pensamiento computacional** es un conjunto de **procesos de pensamiento involucrados en la formulación de problemas y sus soluciones**, de tal manera que las soluciones se representen de una forma que pueda ser llevada a cabo eficazmente por un agente de procesamiento de información.

Formular problemas y soluciones de modo ejecutable por un agente de cómputo

Abstracción, análisis, diseño de representaciones y algoritmos.

Es una **forma en que los humanos, no las computadoras, piensan** para resolver problemas . **No se trata de hacer que los humanos piensen como las computadoras**, sino de usar nuestra inteligencia para aprovechar el poder de la computación.

“Un lenguaje no son solo palabras. Es una cultura, una tradición, una unificación de una comunidad, toda una historia que crea lo que una comunidad es. Todo está contenido en un lenguaje."

Noam Chomsky

Del Caos al Conocimiento: El Poder del Text Mining

Minería de Textos

- Objetivo: descubrir patrones, tendencias y conocimiento en grandes volúmenes de texto.

- Pipeline típico: obtención → limpieza → representación → modelado → evaluación → despliegue/monitoreo

Panorama 2024-2025 en PLN

- Modelos fundamentales y LLMs instruccionales
- Recuperación aumentada (RAG), agentes y herramientas
- Eficiencia: LoRA/QLoRA, cuantización, destilación
- Evaluación: autométrica + humana; robustez y sesgos
- Costos e infraestructura; optimización de inferencia

Nos vemos el jueves en modalidad LAB

Gracias!