Bienvenidos!
Semana 2

Datos No Estructurados y Semiestructurados

Especialización en Economía, opción Ciencia de Datos

FCS - Udelar

Guillermo Lezama

La clase de hoy

- Recomiendo leer: https://paulapereda.com/2025/05/04/datos-que-sean-de-moda-es-realmente-el-marron-en-todas-sus-tonalidades-el-color-de-la-temporada/

- JSON
- XML
- HTML
- Scraping and APIs

- ¿Dudas del cuestionario o alguna de las tareas?

- Comentario sobre las tareas.

XML

- eXtensible Markup Language: un formato de texto para representar datos jerárquicos.
- Usa etiquetas `<etiqueta>` (tags) para marcar inicio y `</etiqueta>` para marcar fin de un elemento.
- También hay atributos `<etiqueta atributo="25">`
- Permite anidar niveles de información.
- <libros>

```
        <libro id="1">
                <titulo>El Principito</titulo>
                <autor>Antoine de Saint-Exupéry</autor>
        </libro>
        <libro id="2">
                <titulo>Cien años de soledad</titulo>
                <autor>Gabriel García Márquez</autor>
        </libro>
</libros>
```

HTML

- HTML (*HyperText Markup Language*) es el lenguaje con el que están construidas la mayoría de las páginas web.

- Actividad con un txt. Este es un párrafo.

- → etiqueta de apertura

- class="intro" → atributo

- "Este es un párrafo." → contenido textual

¿Por qué nos importa en análisis de datos?

- Los datos que necesitamos están en páginas web.
- Podemos usar HTML como fuente de datos no estructurados.
- Podemos extraer texto, títulos, tablas, listas, etc.

En XML:

- Las etiquetas (<ciudad>, <nombre>) definen estructura y contenido de los datos.
- Los atributos (<ciudad nombre="Montevideo">) agregan información adicional dentro de la etiqueta.
- En XML se espera que las etiquetas describan datos y no estilo.

En HTML:

- También usa etiquetas y atributos, pero la mayoría están pensadas para presentación visual (dónde mostrar algo, cómo se ve, qué hace un botón, etc.)
- Por eso hablamos de etiquetas de presentación:
- <h1>, <p>, <b>, <table>, <img>
- Atributos como style=, class=, width=, href=
- En HTML, las etiquetas no describen datos, sino cómo deben mostrarse.

JSON, XML y HTML

- Son las formas comunes de representar datos no estructurados o semiestructurados.
- Son la base para leer, extraer y procesar información desde la web o servicios (como APIs).
- Aparecen en casi todo: desde respuestas de APIs, hasta documentos legislativos o páginas de noticias.

JSON, XML y HTML: Similitudes

- Todos son texto legible por humanos.
- Tienen una estructura jerárquica o anidada.
- Podemos recorrerlos con Python (diccionarios, árboles, selectores).

JSON, XML y HTML: Diferencias

- Estructura (clave → valor, etiquetas y atributos y etiquetas de presentación)
- Json.loads(), ElementTree, BeautifulSoup

API

- API significa *Application Programming Interface*. Es una forma estructurada de pedir datos a un servidor.
- Muchas páginas web y servicios ofrecen APIs para que puedas consultar datos directamente en formato JSON.
- En vez de hacer scraping del HTML, simplemente pedís datos como si fuera una "conversación de robots".

¿Por qué usarlas?

- Devuelven datos limpios (sin etiquetas HTML)
- Más estables y rápidas que el scraping
- Muchas veces pensadas para ser reutilizadas en apps o análisis

- Usamos `requests.get()` para consultar la URL del servicio.