Bienvenidos!

Datos No Estructurados y Semiestructurados

Especialización en Economía, opción Ciencia de Datos

FCS - Udelar

Guillermo Lezama

La clase de hoy

- Intros
- Algunas definiciones
- Una introducción al curso

Bienvenidos

- Introducciones
  - Nombre
  - Que hicieron de sus vidas? (qué estudiaron, dónde trabajaron)
  - Qué hacen de sus vidas? (trabajan, solo estudian, algún buen hobby)
    - Experiencia previa en

      Lingüística / literatura / análisis de textos

      Música / producción o teoría musical

      Artes visuales / fotografía / diseño gráfico

      Comunicación / periodismo / redes sociales

      Programación
  - Que quieren hacer de sus vidas con el diploma?
  - ¿Estás interesado/a en trabajar un proyecto final que esté relacionado con…
  - Una serie que hayas visto, o un libro que hayas leido este año a recomendar (ver o no ver).

Expectativas

- Clases virtuales
  - … pero participen
  - pregunten
    - Las preguntas solo significan que son curiosos
- 10 minutos de cámara.
- Traigan preguntas
- Traigan respuestas
- Traigan problemas

Formalidades

- Trabajos domiciliarios (20%)
  - Formularios múltiple opción (o algo sencillo)
  - Sin nota
  - No todas las clases
  - El objetivo es chequear si entendieron o no.
- Trabajo final (80%)
  - Última clase: presentación en clase (9 de junio)
  - Deadline de trabajo escrito: 28 de Junio
  - En grupo
  - Proyectos y armado de grupos: 29 de mayo

Otras cuestiones para tener en cuenta

- Python
- Libros
  - Voy a seguir varios libros y otros recursos
  - Algunos en el syllabus.
- Ejemplos
- ChatGPT

EVA y horas de oficina

Voy a subir

- Clases grabadas (en el correr del dia y cercano a la noche)
  - No todo el contenido
- Google Colab (subidos antes de comenzar la clase)

Disponible para reunirme entre semana por zoom

Escriban a guillermo.lezama@cienciassociales.edu.uy

Qué son datos no estructurados?

- Datos que carecen de estructura 🤡
- Entonces, qué son datos estructurados?
- Información organizada en una forma pre-definida y consistente
- Columnas y filas fijas (un excel)
- $P(votar\ a\ alcalde)_i = \alpha + \beta_1 Educación_i + \beta_2 votarFA_i + \\ + \beta_3 edad_i + \ \varepsilon_i$
- Necesitamos una tabla

Otras ejemplos de base de datos estructuradas

- Base de datos relacionales
  - FCS tiene un dataset con sus estudiantes
  - Cada curso tiene un dataset con sus notas
  - Hay un dataset con todas las inscripciones de este semester
  - Cada curso tiene un programa
  - Todas relacionadas con su cédula, código de curso
- Datos de panel
  - Tiempo e individuos
- Datos de mapas
  - Latitudes, longitudes y otra variable

Que información carece de estructura?

1) Semi-estructurados

🎓 Formulario:

- Nombre: Mariana
- Edad: 25
- Intereses: Cine, lectura, programación
- Comentarios adicionales: Me interesa aplicar esto a análisis de redes sociales.

```
{
"nombre": "Mariana",
"edad": 25,
"intereses": ["Cine", "lectura", "programación"],
"comentarios": "Me interesa aplicar esto a análisis de redes sociales."
}
{
  "nombre": "Luis",
  "comentarios": "Prefiero no decir mi edad."
}
```


- Diccionario -> JSON (datos electorales)
- RSS feeds de noticias
- HTML, XML
- Historias clínicas

Que información carece de estructura?

2) No estructurados

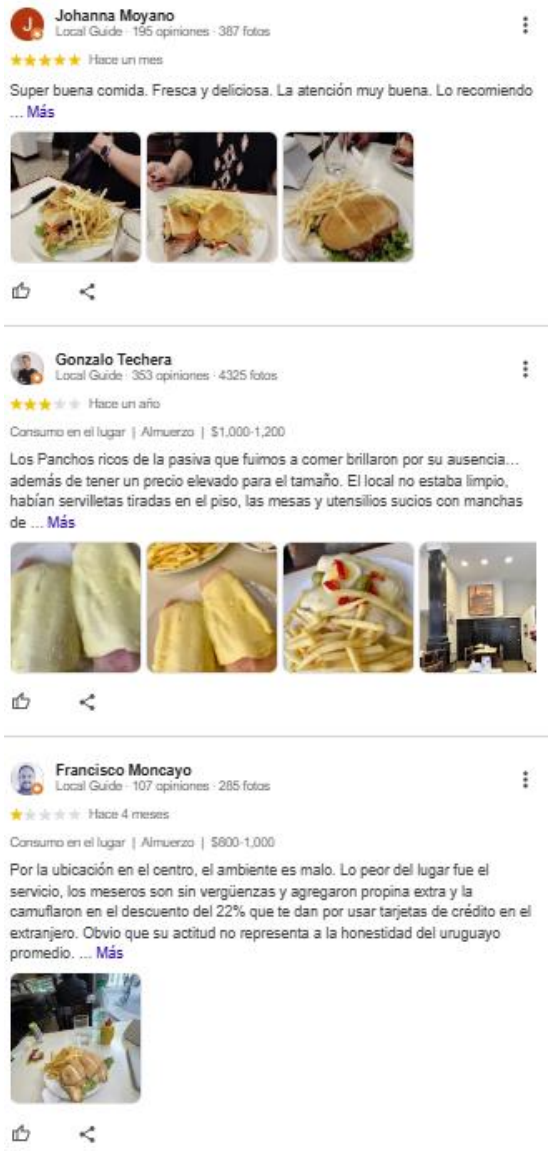
📝 **Texto libre de la reseña**

No sigue ningún formato. Podría tener emociones, sarcasmo, menciones específicas, etc.

🖼️ **Fotos subidas por el usuario**

Contienen información visual que se puede analizar (tipo de comida, limpieza, presentación), pero no están etiquetadas automáticamente.

Otros ejemplos: Audio, video, pdfs



Objetivo del curso

- Algunos datos en bruto no se pueden analizar directamente
- No puedo hacer una regresión con fotos de panchos
- Extraer información del caos.
- Pero qué?
    1) Para qué? Qué nos interesa?
    2) Herramientas que ya existen para transformar caos en datos
    3) Automatizar!

Pasar de un archivo desordenado a un dataset limpio.

Procedimientos Teóricos

- La pregunta!
- Análisis de Datos: Realizar un análisis exhaustivo de los datos no estructurados para comprender su contenido.
- Identificación de Patrones
- Diseño de Esquema: Definir un esquema de datos adecuado que pueda capturar la estructura necesaria para representar los datos de manera organizada.
- Selección de Herramientas y Tecnologías: Seleccionar las herramientas y tecnologías adecuadas para realizar la conversión y el procesamiento de los datos.
- Implementarlo

Estructura del Curso

1) Formatos comunes: **JSON, XML,** YAML, HTML (~ 2 clases)

2) Procesamiento de texto (3 clases)

3) Procesamiento de imágenes y sonidos, ¿ubicación? ¿datos deportivos? (2 clases)

4) Extraer data caótica (Scraping, APIs) (1 clase)

5) Otros ejemplos (1 clase)

Lectura, exploración y estructuración.

Hoy

- Introducción a Python
- Diccionarios y JSON
- XML
- No requieren un esquema rígido definido de antemano, pero contienen un esquema implícito (keys/etiquetas) que puede inferirse al leer los datos.

- En algún momento, actividad grupal (en EVA).

Qué es Python, cómo funciona y más en 1 slide

- Un lenguaje sencillo que te permite indicarle a la computadora qué hacer.

- 1. Escribís tus instrucciones (código).
- 2. Python las lee y las ejecuta paso a paso.

- Listas
- - Una fila donde guardas varios objetos en orden.
- - Ejemplo: mi_lista = ["manzana", "banana", "cereza"]

- Mostrar un resultado en pantalla: print(mi_lista)

Qué es una library?

- Es una colección de herramientas (funciones, clases y datos) creadas por otros.

- Ejemplo de la vida real: Calculadora vs. hacer las cuentas manuals

- pandas
- numpy
- requests
- xml.etree.ElementTree

Pandas 🐼

- - Una library de Python que convierte tus listas y diccionarios en tablas

- Es como una hoja de Excel dentro de tu código: podés filtrar, ordenar y resumir datos fácilmente.

- Estructura principal: DataFrame
  - Una tabla con filas (cada registro) y columnas (cada variable).

- import pandas as pd

Diccionarios

- Python data structure
- Es un objeto (no plain text)
- key a values

```
{
   “nombre": "Ana",
   “edad": 30,
   “materias favoritas": [“datos no estructurados", “estadistica"]
}
```

```
[{
   “nombre": "Ana",
   “edad": 30,
   “materias favoritas": [“datos no estructurados", “estadistica"]
},
{
   "nombre": "Luis",
   "edad": “veinte”,
   "materias favoritas": ["machine learning", “analisis exploratorio”]
 }]
```

Por qué son semi-estructurados?

- Flexibilidad de keys (cada entrada con diferentes keys)
- Valores heterogéneos (misma key con distintos formatos)
- No tiene un formato tabular estricto

JSON

- JavaScript Object Notation
- Almacenar y transmitir datos, sobretodo cuando descargamos información
- Es el idioma de las APIs y muchas base de datos
- Es más flexible que una tabla
- Es un formato de texto
- Se usa para almacenar listas, diccionarios, strings, etc.
- El uso más común, es guardar diccionarios

XML

- eXtensible Markup Language: un formato de texto para representar datos jerárquicos.
- Usa etiquetas `<etiqueta>` (tags) para marcar inicio y `</etiqueta>` para marcar fin.
- 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>
```