

# MÓDULO 01: CARGA Y VALIDACIÓN DE TRANSCRIPCIONES

```
In [1]: # -----
# ATENCION - FIJAR ESTAS VARIABLES ANTES DE EJECUTAR
# -----

nombre_lote = "LOTE_20250614"

nombre_modulo = "MODULO_01"

# -----
# Configuración del entorno (Colab y Local)
# -----

try:
    import google.colab
    EN_COLAB = True
except ImportError:
    EN_COLAB = False

if EN_COLAB:
    from google.colab import drive
    drive.mount("/content/drive", force_remount=True)
    ruta_base = "/content/drive/MyDrive/TFM_EVA_MARTIN/Modulos"
else:
    ruta_base = "G:/Mi unidad/TFM_EVA_MARTIN/Modulos"

print(f"Entorno detectado: {'Google Colab' if EN_COLAB else 'Local'}")
print(f"Ruta base: {ruta_base}")
```

Entorno detectado: Local

Ruta base: G:/Mi unidad/TFM\_EVA\_MARTIN/Modulos

```
In [2]: # -----
# Cargar utilidades comunes
# -----

import os
import sys
import pandas as pd
ruta_config = os.path.join(ruta_base, "config.yaml")

if ruta_base not in sys.path:
    sys.path.append(ruta_base)
import yaml
import utilidades_comunes
```

```
In [3]: # -----
# PASO 1: Configurar Logger
# -----
# 1. Configurar Logger
logger = utilidades_comunes.configurar_logger(nombre_modulo, ruta_logs=os.path.j

# 2. Inicializar entorno
entorno = utilidades_comunes.inicializar_entorno(nombre_modulo, nombre_lote, rut
```

```

ruta_entrada = os.path.join(os.path.dirname(ruta_base), "DATA", nombre_lote)

# Este módulo es el primero del flujo, por tanto se sobrescribe la ruta_entrada
entorno["ruta_entrada"] = os.path.join(os.path.dirname(ruta_base), "DATA", nombre_lote)
logger.info(f"⚠ MODULO_01: se ha sobrescrito entorno['ruta_entrada'] con {entorno['ruta_entrada']}")

```

```

2025-06-15 15:17:42,284 - INFO - 📁 Entorno inicializado para MODULO_01
2025-06-15 15:17:42,287 - INFO - 📁 Ruta entrada: G:/Mi unidad/TFM_EVA_MARTIN/Modulos\MODULO_00\./salida
2025-06-15 15:17:42,287 - INFO - 📁 Ruta salida: G:/Mi unidad/TFM_EVA_MARTIN/Modulos\MODULO_01\./salida
2025-06-15 15:17:42,287 - INFO - 📁 Ruta logs: G:/Mi unidad/TFM_EVA_MARTIN/Modulos\MODULO_01\./logs
2025-06-15 15:17:42,304 - INFO - 📁 Ruta ejemplos: G:/Mi unidad/TFM_EVA_MARTIN/Modulos\MODULO_01\./ejemplos
2025-06-15 15:17:42,306 - INFO - 🔗 Módulo anterior: MODULO_00
2025-06-15 15:17:42,308 - INFO - 🆔 Lote ID: 20250614
2025-06-15 15:17:42,311 - INFO - ⚠ MODULO_01: se ha sobrescrito entorno['ruta_entrada'] con G:/Mi unidad/TFM_EVA_MARTIN\DATA\LOTE_20250614

```

```

In [4]: # -----
# PASO 2: Cargar dataset entrada (conjunto de .txt)
# -----
@utilidades_comunes.medir_tiempo
def cargar_transcripciones_txt(ruta_entrada):
    """
    Recorre la carpeta de .txt y devuelve un DataFrame con:
    - nomfichero: nombre del fichero
    - Etiqueta: parte del nombre antes del primer '_' (o adaptar extracción)
    - Texto: contenido del fichero
    """
    registros = []
    for nomfichero in os.listdir(ruta_entrada):
        if not nomfichero.endswith('.txt'):
            continue
        nomfichero_path = os.path.join(ruta_entrada, nomfichero)
        try:
            with open(nomfichero_path, 'r', encoding='utf-8') as f:
                transcripcion = f.read().strip()
        except UnicodeDecodeError:
            with open(nomfichero_path, 'r', encoding='latin-1') as f:
                transcripcion = f.read().strip()
        if transcripcion:
            # Extraer etiqueta; por ejemplo, parte antes del primer guión bajo
            etiqueta = nomfichero.replace('.txt', '').split('_')[0]
            registros.append({
                'nomfichero': nomfichero,
                'etiqueta': etiqueta,
                'transcripcion': transcripcion
            })
    df = pd.DataFrame(registros, columns=['nomfichero', 'etiqueta', 'transcripcion'])
    return df

df_entrada = cargar_transcripciones_txt(ruta_entrada)

```

🕒 'cargar\_transcripciones\_txt' completado en 0.07 s.

```

In [5]: # -----
# PASO 3: Procesamiento específico del módulo

```

```
# En este caso, simplemente eliminamos duplicados como validación mínima
# -----
def procesamiento_modulo_01(df):
    registros_antes = len(df)
    df = df.drop_duplicates()
    registros_despues = len(df)
    logger.info(f"Duplicados eliminados: {registros_antes - registros_despues}")
    return df

df_salida = procesamiento_modulo_01(df_entrada)
```

2025-06-15 15:17:51,291 - INFO - Duplicados eliminados: 0

```
In [6]: # -----
# PASO 4: Validación post-procesamiento
# -----

utilidades_comunes.validar_integridad(df_salida, logger)
```

2025-06-15 15:17:54,726 - INFO - 🔍 Validando integridad del dataset...

2025-06-15 15:17:54,729 - INFO - ✅ Columnas requeridas presentes: ['nomfichero', 'etiqueta', 'transcripcion']

2025-06-15 15:17:54,732 - INFO - ⚠️ Dimensiones del dataset: 3 filas, 3 columnas

2025-06-15 15:17:54,736 - INFO - ✅ Validación de integridad completada correctamente.

```
In [7]: # -----
# PASO 5: Guardar dataset salida con nombre estándar
# -----

nombre_salida = os.path.join(
    entorno["ruta_salida"],
    f"dataset_{nombre_modulo.lower()}_{entorno['lote_id']}.csv"
)

utilidades_comunes.guardar_dataset(df_salida, nombre_salida, logger=logger)
```

2025-06-15 15:17:58,734 - INFO - 📁 Dataset guardado en: G:/Mi unidad/TFM\_EVA\_MARTIN/Modulos\MODULO\_01\./salida\dataset\_modulo\_01\_20250614.csv (3 filas, 3 columnas)



```
In [8]: # -----
# PASO 6: Mostrar muestra final
# -----

nombre_muestra = f"{nombre_modulo.lower()}_{entorno['lote_id']}"

utilidades_comunes.mostrar_muestra_dataset(df_salida, nombre_muestra, logger=logger)


logger.info(f"✅ Finalización del procesamiento del {nombre_modulo}")
logger.info(f"📁 Dataset final disponible en: {nombre_salida}")
```

```

2025-06-15 15:18:01,760 - INFO - --- Muestra de modulo_01_20250614 (primeras 5 fi
las) ---
2025-06-15 15:18:01,763 - INFO - Filas totales: 3, Columnas totales: 3
2025-06-15 15:18:01,789 - INFO -
| nomfichero
| etiqueta | transcripcion
|
|:-----|:-----|:-----|:-----|
|-----|:-----|:-----|:-----|
|-----|:-----|:-----|:-----|
|-----|:-----|:-----|:-----|
|-----|:-----|:-----|:-----|
|-----|:-----|:-----|:-----|
|-----|:-----|:-----|:-----|
|-----|:-----|:-----|:-----|
| Positivo_[965021512]_MERIDIANO_900_GEN._2023-02-27_20-57-01.txt
| Positivo | Sí, sí, soy XXXXXX. Primero que todo, queríamos transmitir nuestra
s más sinceras condolencias, señora XXXXXX. Muchas gracias. Llegamos simplemente
para confirmar que tanto el servicio como la atención que estaban recibiendo uste
des los familiares eran ambos correctos. Sí. Muchas gracias. Hasta luego.
|
| Neutro_SD23-03516_[965021512]_MERIDIANO_900_GEN._2023-02-27_10-18-51_674114825.
txt | Neutro | Bienvenido Meridiano. Por su seguridad, esta llamada podrá ser
grabada. Usted consiente en que los datos que facilite se incorporen en un fichero
o titularidad de Meridiano S.A. con la finalidad de gestionar la prestación del s
ervicio. Si ya conoce nuestra política de protección de datos y no desea escuchar
la nuevamente, pulse o diga cero. En breve será atendido por uno de nuestros agen
tes. Por favor, no se retire. |
| Positivo_[965021512]_MERIDIANO_900_GEN._2023-02-28_15-57-35.txt
| Positivo | de asistencia. En primer lugar, nuestras más sinceras condolencias
por el fallecimiento de su padre, don XXXXXX. El motivo de mi llamada era comprob
ar si el servicio de su padre se había desarrollado correctamente, si había algo
que pudiéramos hacer para asistirles. ¿Tenían alguna duda del proceso? No, de mom
ento va bien, la asistencia va bien. De acuerdo. Muy bien, pues muchas gracias.
|
2025-06-15 15:18:01,804 - INFO -
--- Estadísticas básicas ---
| count | unique | top
| freq |
|:-----|:-----|:-----|:-----|
|-----|:-----|:-----|:-----|
|-----|:-----|:-----|:-----|
|-----|:-----|:-----|:-----|
|-----|:-----|:-----|:-----|
| nomfichero | 3 | 3 | Positivo_[965021512]_MERIDIANO_900_GEN._20
23-02-27_20-57-01.txt
| 1 |
| etiqueta | 3 | 2 | Positivo
| 2 |
| transcripcion | 3 | 3 | Sí, sí, soy XXXXXX. Primero que todo, quer
íamos transmitir nuestras más sinceras condolencias, señora XXXXXX. Muchas gracia
s. Llegamos simplemente para confirmar que tanto el servicio como la atención que
estaban recibiendo ustedes los familiares eran ambos correctos. Sí. Muchas gracia
s. Hasta luego. | 1 |
2025-06-15 15:18:01,815 - INFO - -----
2025-06-15 15:18:01,818 - INFO -  Finalización del procesamiento del MODULO_01
2025-06-15 15:18:01,825 - INFO -  Dataset final disponible en: G:/Mi unidad/TF
M_EVA_MARTIN/Modulos\MODULO_01\./salida\dataset_modulo_01_20250614.csv

```

```
In [9]: # -----  
# PASO 7: Guardar muestra en carpeta /ejemplos  
# -----  
  
utilidades_comunes.guardar_muestra_dataset(  
    df_salida,  
    nombre_muestra,  
    entorno["ruta_ejemplos"],  
    logger=logger  
)
```

2025-06-15 15:18:12,240 - INFO -  Muestra guardada en G:/Mi unidad/TFM\_EVA\_MARTIN/Modulos\MODULO\_01\./ejemplos\muestra\_modulo\_01\_20250614.csv (3 filas)