

⌚ Pipeline de Datos – Entregable 1: Análisis y Visualización

Deadline: 30/06

Este pipeline representa el flujo de trabajo lógico (workflow) del Entregable 1, el cual servirá como cimiento para el scraping masivo, la curación, y el modelado posterior.



Made with  Napkin

◊ 1. Ingesta de Datos Inicial

- **Fuente:** Dataset descargado de [Argendir](#) y <https://tranco-list.eu/> + dataset de datos sintéticos
 - **Objetivo:** Contar con un primer set estructurado de sitios web con sus metadatos (categoría, título, descripción, etc.).
 - **Tareas:**
 - Validar estructura del archivo.
 - Analizar codificación, delimitadores, y limpieza básica de strings.
 - Eliminar filas incompletas, duplicadas o irrelevantes.
-

◊ 2. Preprocesamiento Estructural

- **Objetivo:** Generar una versión funcional del dataset para análisis exploratorio.
 - **Tareas:**
 - Normalización de columnas (nombres, tipos).
 - Extracción y estandarización de categorías temáticas.
 - Creación de campos derivados: dominio base, TLD (.com.ar, .org, etc.), longitud del texto, keywords.
-

◊ 3. Análisis Exploratorio de Datos (EDA)

- **Objetivo:** Comprender las características principales de los sitios listados.
 - **Tareas:**
 - Distribución de sitios por categoría.
 - Frecuencia de aparición por dominio y TLD.
 - Nube de palabras de títulos y descripciones.
 - Visualización de la cobertura de categorías (mapa de calor o treemap).
 - Detección de outliers en nombres sospechosos (por regex, términos comunes en fraudes, etc.).
-

◊ 4. Scraping Complementario Inicial

- **Objetivo:** Enriquecer algunos registros con métricas externas (opcional en esta fase si no hay tráfico disponible en el CSV).
- **Posibles tareas:**
 - Obtener estimaciones de tráfico (via SimilarWeb, Alexa, etc.).

- Validar accesibilidad de los sitios: HTTP status, redirecciones, certificados SSL.
 - Extraer títulos HTML, metatags, favicons, etc.
-

◊ 5. Visualización

- **Objetivo:** Comunicar hallazgos visuales y patrones iniciales.
 - **Tareas:**
 - Histogramas de categorías más frecuentes.
 - Gráfico de barras de dominios más comunes.
 - Wordclouds de descripciones.
 - Mapa de árbol de categorías + subcategorías.
 - Gráfico de líneas temporales si hay fechas de actualización o scraping.
-

◊ 6. Data Insights Preliminares

- ¿Qué categorías predominan?
 - ¿Existen clusters sospechosos (por ejemplo, muchos sitios con palabras como "premio", "banco", "gratuito")?
 - ¿Hay señales tempranas de duplicación o generación automática de sitios?
 - ¿Qué dominios podrían ser priorizados para scraping o verificación de legitimidad?
-

◊ 7. Documentación y Output

- **Notebook Jupyter o Google Colab** con:
 - Limpieza y análisis comentados.
 - Visualizaciones interactivas o exportadas.
- **Documento técnico** con:
 - Hipótesis iniciales.
 - Metodología del análisis.
 - Hallazgos.
 - Recomendaciones para el Entregable 2.
- **Dataset intermedio limpio** (argendir_clean.csv).

Anexo con características que debe cumplir el entregable:

- Los proyectos de Ciencia de datos, son proyectos que dependen mucho del “ojo” de cada observador, por lo cual es recomendable que estén fuertemente estructurados. Se recomienda usar la PEP8, (guía que indica las convenciones estilísticas a seguir para escribir código <https://ellibrodepython.com/python-pep8>)
- Desarrollar cada punto en la misma notebook donde se escriba el código. Dicho notebook debe contar con un índice, con sus diferentes apartados y el código debe ser fácil de leer, estar probado y comentado (esto último, en función de la necesidad).
- Enviar el link directo del archivo .ipynb ó alternativamente subir el entregable a un repositorio GitHub mediante la integración con Google Colab. Recordar que al compartir el notebook, queden habilitados los permisos de edición, para poder dejar comentarios/correcciones. Tener en cuenta que si bien, se pueden realizar diversos análisis y visualizaciones, se debe dejar en el entregable sólo aquello que sea relevante.

Repositorio:

https://github.com/NoeliaFerrero/Proyecto_Mentoria_FAMAF_2025.git

¡Buen comienzo en su primer Proyecto de Data Science!