Proyecto Final: Desarrollo de un Sistema de Análisis de Noticias Salvadoreñas

Objetivo General:

El objetivo principal de este proyecto es aplicar los fundamentos de ciencia de datos para desarrollar un sistema integral de análisis de noticias salvadoreñas. Se abordarán diferentes etapas, desde la obtención de datos mediante web scraping hasta la implementación de modelos de aprendizaje automático y procesamiento del lenguaje natural, culminando con la creación de una interfaz interactiva utilizando Streamlit.

Tareas Específicas:

1. Recolección de datos:
    - Utilizar técnicas de web scraping o APIs para obtener noticias de un periódico salvadoreño, por ejemplo El Diario de Hoy

2. EDA y Visualización de Datos:
    - Realizar un Análisis Exploratorio de Datos (EDA) en las noticias recolectadas.
    - Agrupar las noticias por fecha y categoría.
    - Visualizar de manera efectiva los patrones y tendencias encontrados en los datos.

3. Aprendizaje Automático (ML):
- Construir un modelo de clasificación supervisada multietiqueta que, dados los contenidos de las noticias, genere la categoría correspondiente.
- Evaluar y validar el rendimiento del modelo utilizando métricas pertinentes.

4. Procesamiento del Lenguaje Natural (NLP):
- Implementar técnicas de procesamiento del lenguaje natural para realizar resúmenes de texto y/o extracción de palabras clave para cada noticia o categoría.

5. Streamlit:
- Mostrar las visualizaciones generadas en las etapas anteriores utilizando Streamlit.
- Desarrollar una página interactiva en Streamlit que permita a los usuarios pegar un texto de noticias y obtener sus categorías, palabras clave o resumen utilizando los modelos construidos en las etapas anteriores.

6. GitHub y DagsHub:
- Subir todos los archivos y código del proyecto a DagsHub.
- Generar un archivo README que describa de manera detallada la estructura del proyecto, los pasos para replicar los resultados y cualquier información relevante.

Entregables:

- Código fuente del proyecto alojado en DagsHub.
- Archivo README en el repositorio de DagsHub.
- Página interactiva en Streamlit que demuestre la funcionalidad del sistema.
- Informe técnico detallado que describa la metodología, resultados y conclusiones del proyecto.

Presentación:

- Preparar una presentación en video, que puede realizarse durante una semana especial de presentaciones. Cada grupo tendrá entre 10 y 15 minutos para mostrar y explicar su trabajo.

Fecha de Entrega:

- La primer entrega (partes 1 y 2, Webscraping y EDA) es el **8/12/2024**
- 5 de enero: Machine Learning
- 12 de enero: NLP
- La fecha de entrega del proyecto completo es el **19/01/2024**

Este proyecto permitirá a los estudiantes aplicar de manera práctica los conceptos aprendidos en el curso de Fundamentos de Ciencia de Datos y desarrollar habilidades en el manejo de datos, aprendizaje automático, procesamiento del lenguaje natural y creación de interfaces interactivas.