Alcanzando los objetivos de desarrollo sostenible. Un aporte desde el machine learning

La Organización de las Naciones Unidas (ONU) adopta, el 25 de septiembre del año 2015, la Agenda 2030 para el desarrollo sostenible, cuyo fin es reducir la pobreza, garantizar acceso a la salud y educación, buscar igualdad de género y oportunidades y disminuir el impacto ambiental, entre otros propósitos. A tal efecto, define 17 objetivos de desarrollo sostenible (ODS) y 169 metas (derivadas de los diferentes ODS). Dentro del trabajo en conjunto de diferentes entes para alcanzar el cumplimiento de los objetivos de desarrollo sostenible definidos en la Agenda 2030, muchas entidades tienen como función el seguimiento y la evaluación de las políticas públicas y su impacto a nivel social. Este es el caso del Fondo de Población de las Naciones Unidas (UNFPA, por sus siglas en inglés), que, junto con instituciones públicas y haciendo uso de diferentes herramientas de participación ciudadana, busca identificar problemas y evaluar soluciones actuales, relacionando la información con los diferentes ODS.

Uno de los procesos que requiere de un mayor esfuerzo es la interpretación y análisis de la información textual procedente de diferentes fuentes implicadas en la planeación participativa para el desarrollo a nivel territorial, ya que es una tarea que consume gran cantidad recursos y para la cual se requiere de expertos que relacionen los textos con los ODS. Este conocimiento facilitaría la toma de decisiones con base en la opinión de la población y permitiría encaminar las políticas públicas de manera más informada para el cumplimiento de la Agenda 2030.

El objetivo del proyecto es desarrollar una solución, basada en técnicas de procesamiento del lenguaje natural y machine learning, que permita clasificar automáticamente un texto según los 17 ODS, ofreciendo una forma de presentación de resultados a través de una herramienta de fácil comprensión para el usuario final.

A. Objetivo.

- Desarrollar una solución, basada en técnicas de procesamiento de lenguaje natural y machine learning, que facilite la interpretación y análisis de información textual para la identificación de relaciones semánticas con los Objetivos de Desarrollo Sostenibles.

B. Conjunto de datos.

El conjunto de datos forma parte del proyecto "OSDG Community Dataset"[1] (OSDG-CD), en su versión 2023, que contiene un total de 40.067 textos, de los cuales 3000 provienen de fuentes relacionadas con las Naciones Unidas. También contiene documentos públicos, resúmenes de artículos y reportes. La plataforma reúne investigadores, expertos en la materia y defensores de los ODS de todo el mundo para crear una fuente amplia y precisa de información textual sobre los ODS. Los voluntarios de la comunidad utilizan la plataforma para participar en ejercicios de etiquetado en los que validan la relevancia de cada texto para los ODS basándose en sus conocimientos previos.

Los textos que serán utilizados en este proyecto han sido traducidos al español por herramientas como Deepl[2]. También se realizó una aumentación de textos a través del API de ChatGPT[3].

C. Actividades para realizar.

1. Preparación de los textos utilizando el esquema de bolsa de palabras (BOW)l con una pesado TF-IDF. Para este paso construir un pipeline que integre las transformaciones que se consideren adecuadas.
2. Desarrollo de un modelo de clasificación que permita relacionar un texto con un ODS. Para manejar la complejidad del espacio de entrada aplicar un algoritmo de reducción de la dimensionalidad
3. Evaluación del modelo con textos que no hayan sido utilizados para el aprendizaje.

D. Consideraciones.

El algoritmo de clasificación a utilizar, así como la técnica de reducción de la dimensionalidad, queda a consideración de cada grupo, pero es importante justificar la elección.

Es posible utilizar otra técnica para generar la representación vectorial de los textos, como Word2Vec.

E. Entregable.

Notebook (*.ipynb y *.html) del método desarrollado. El Notebook debe estar documentado con las justificaciones de las decisiones tomadas en cada paso. Además, deben ser visibles las ejecuciones de cada celda. Para evidenciar el desempeño del método construido el notebook debe mostrar las clasificaciones para al menos cuatro textos del conjunto test.

[1] OSDG Community Dataset (OSDG-CD). (https://osdg.ai/news/New-release-of-OSDG-Community-dataset).
[2] https://www.deepl.com/es/translator
[3] https://chat.openai.com/g/g-I1XNbsyDK-api-docs

Esta entrega debe realizarse al final de la semana 7, en donde encontrarás un espacio para adjuntar los dos archivos.

F. Criterios de evaluación.

Actividad	Porcentaje	Objetivos de aprendizaje
Preparación de los datos, incluyendo la reducción de la dimensionalidad, justificando las decisiones tomadas.	35%	Construir representaciones vectoriales de documentos para su análisis por algoritmos de aprendizaje. Aplicar técnicas de reducción de la dimensionalidad sobre textos para resolver problemas de aprendizaje y visualización. Caracterizar las fases del proceso de desarrollo de un proyecto de machine learning. Identificar las diferencias entre los métodos de reducción de la dimensionalidad por selección y por transformación.
Construcción del pipeline de preparación de datos.	15%	Reconocer la importancia de la preparación de los datos como etapa previa a la aplicación de los algoritmos de aprendizaje.
Construcción del modelo de clasificación con el algoritmo seleccionado con búsqueda de hiperparámetros, validándolo con medidas de evaluación adecuadas. Se justifica la selección del algoritmo y las métricas empleadas.	35%	Caracterizar las fases del proceso de desarrollo de un proyecto de machine learning. Conocer cómo las técnicas de reducción de la dimensionalidad pueden ser utilizadas en el análisis de textos.
Evidencia del desempeño del método construido mostrando	15%	Caracterizar las fases del proceso de desarrollo de un proyecto de machine learning. Conocer cómo las técnicas de reducción de la dimensionalidad pueden ser utilizadas en el análisis de textos.