Karina Alejandra Valladares Díaz (18005)
José Eduardo López Gómez (181045)
Alexa Carolina Bravo Vicente (18831)

Proyecto 2
Procesamiento del lenguaje natural con tweets de desastres naturales

Problema científico

El procesamiento de lenguaje natural, o NLP, por sus siglas en inglés, es una rama de la ciencia de computación enfocada en otorgar a las computadoras la habilidad de comprender y entender palabras habladas y escritas, de forma muy similar a la que los humanos lo hacemos. Para ello, se combina el modelaje del lenguaje humano con procesos estadísticos, *machine learning* y *deep learning*. Dada la ambigüedad del lenguaje humano, existen múltiples técnicas y prácticas dentro del NLP, con el fin de volverlo más sencillo de procesar para las máquinas (IBM Cloud Education, 2020).

Dada la complejidad del lenguaje humano, en general existen técnicas preliminares para desglosar el lenguaje, de forma que colabora con la computadora en la comprensión de lo que está analizando. Entre estas se incluyen:

- Reconocimiento de habla: es la habilidad en la que el texto hablado se convierte en un texto escrito. Dicha tarea es necesaria para cualquier aplicación que sigue con comandos de voz o responde a preguntas orales.
- Clasificación de categorías gramaticales: en este proceso se determinan qué categorías gramaticales están presentes dentro de cada oración y se van clasificando de acuerdo a estas.
- Desambiguación del sentido de las palabras: es la selección del sentido y significado de las palabras con varios significados a partir de un análisis semántico.
- Análisis sentimental: intenta la extracción de cualidades subjetivas del texto.

Objetivos

El objetivo principal del proyecto es la creación de un modelo de aprendizaje automático que sea capaz de analizar y evaluar los datos presentados en *tweets*, para determinar si la información presentada en ellos es verdadera o falsa. Para ello, vamos a dividir los objetivos en porciones más pequeñas.

Los objetivos específicos para el proyecto son:

- Definir un problema científico real y determinar un método de solución para el mismo.
- Determinar qué técnicas usadas dentro del Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés) son pertinentes para el estudio en cuestión.
- Realizar un análisis detallado del problema planteado y crear una solución teórica para el mismo.
- Realizar un proceso de limpieza para los datos, puesto que son lenguaje natural.

Análisis exploratorio inicial

Para este proyecto se utilizará únicamente el conjunto de datos de entrenamiento *train.csv*. Este dataset cuenta con 8544 datos los cuales se dividen en cinco variables:

Variable	Tipo	Descripción
id	Numérica	Identificador de cada tweet
keyword	Categórica	Palabra clave del tweet
location	Categórica	Ubicación de donde se envió el tweet
text	Categórica	Texto del tweet
target	Numérica	1 si el tweet es de un caso real, 0 en caso contrario

Se realizará una limpieza de datos:

- Convertir todo a minúsculas
- Eliminación de números o caracteres especiales
- Eliminación de referencias a externos, como "@user" o hilos de información
- Eliminación de emojis y signos de puntuación

Se buscará la frase más usada, las palabras más repetidas, se hará una nube de palabras con estas y gráficos qué expliquen las variables.

Referencias consultadas

IBM Cloud Education. (2020). *Natural Language Processing (NLP).* IBM Cloud Learn Hub. Recuperado de: https://www.ibm.com/cloud/learn/natural-language-processing