FUNDACIÓN TECNOLÓGICA DE COSTA RICA
PROGRAMA CIENCIA DE DATOS
CURSO PROCESAMIENTO DE LENGUAJE NATURAL TAREA 5

Twitter se ha convertido en un canal de comunicación importante en tiempos de emergencia. La ubicuidad de los teléfonos inteligentes permite a las personas anunciar una emergencia que están observando en tiempo real. Debido a esto, cada vez hay más agencias interesadas en monitorear Twitter de manera automática (es decir, organizaciones de ayuda humanitaria y agencias de noticias).

Sin embargo, no siempre está claro si las palabras de una persona realmente están anunciando un desastre. Por ejemplo, si alguien escribe:

"Por el lado positivo, MIRA EL CIELO AL ATARDECER, ¡ESTABA EN LLAMAS!"

El autor utiliza explícitamente la frase "EN LLAMAS", pero lo dice metafóricamente. Esto resulta evidente para cualquier persona de inmediato, especialmente si viene acompañado de una imagen.

Se deben construir y comparar los resultados obtenidos a partir de dos modelo de aprendizaje automático que predigan qué *tweets* tratan sobre desastres reales y cuáles no. Para esos modelos se usará ***Support Vector Machines*** (SVM) y Redes Recurrentes *Long Short-Term Memory* (LSTM). Se dispondrá de un conjunto de datos de 10 000 tweets que fueron clasificados manualmente.

Cada registro del conjunto de datos tiene las siguientes columnas con la información descrita:

- **id**: un identificador único para cada tweet
- **text**: el texto del tweet
- **location**: la ubicación desde la que se envió el tweet (puede estar en blanco)
- **keyword**: una palabra clave particular del tweet (puede estar en blanco)
- **target**: indica si un tweet trata sobre un desastre real (1) o no (0)

A) *Support Vector Machines* (SVM)

Escribir un cuaderno de Jupyter que realice las siguientes acciones. Puede basarse en el cuaderno sobre SVM visto en clase

Semana 05 Clasificación de texto – SVM

Debe asegurarse de que el cuaderno acceda al archivo de datos usando un path que no dependa de su máquina.

Carga (10 puntos)

1. Cargar el archivo '**tweets.csv**' que se distribuye con este enunciado.
2. Contar la frecuencia de las dos clases y calcular el porcentaje de cada clase.
3. Obtener la lista de stopwords para inglés.
4. Calcular y mostrar la distribución de palabras para cada una de las clases. Antes de contar debe eliminar las palabras no útiles (stopwords)

Pre-procesamiento (5 puntos)

5. Separar la colección en un conjunto de entrenamiento y uno de prueba.
   Dar un 80% de los registros al conjunto de entrenamiento.

6. Convertir el conjunto de entrenamiento y el conjunto de pruebas a la representación requerida por el modelo SVM.

Experimentos (10 puntos)

7. Implementar y entrenar un clasificador usando SVM.
8. Obtener predicciones del modelo SVM usando el conjunto de prueba.

Evaluación (5 puntos)

9. Obtener los valores de precisión, recall, f1-score y acierto para el modelo. Tanto globales como por clase.
10. Comentar los resultados obtenidos.

B) *Long Short-Term Memory* (LSTM)

Pre-procesamiento (5 puntos)

11. Utilice las mismas particiones de entrenamiento y pruebas obtenidas en el ejercicio A)
12. Convertir el conjunto de entrenamiento y el conjunto de pruebas a la representación requerida por el modelo LSTM.

Experimentos (10 puntos)

13. Utilice la biblioteca Pytorch para implementar y entrenar un clasificador usando una red recurrente LSTM.
14. Durante el entrenamiento grafique la curva de error, explique los resultados obtenidos en gráfica y ajuste el modelo o el proceso de entrenamiento apropiadamente (por ejemplo, verifique que el modelo no esté sobre-ajustado).
15. Obtener predicciones del modelo LSTM usando el conjunto de prueba.

Evaluación (5 puntos)

16. Obtener los valores de precisión, recall, f1-score y acierto para el modelo. Tanto globales como por clase.
17. Comente y compare los resultados obtenidos con SVM y LSTM.