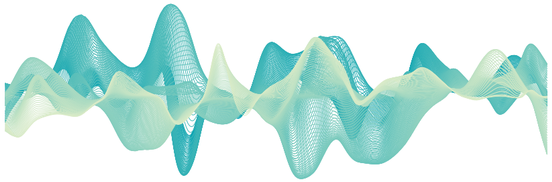
RECONOCIMIENTO DE AUDIO USANDO INTELIGENCIA ARTIFICIAL

KEVIN RINCON - LISI

CONTENIDO

AUDIO



Ondas

Infinitas

Natural

MUESTREO

Puntos

Finitos

Digital

MUESTREO

El **muestreo** es el proceso de medir el valor de una señal continua en pasos de tiempo fijos.

TASA DE MUESTREO

La **tasa de muestreo** es el número de muestras tomadas en un segundo y se mide en hercios (Hz).

REMUESTREO

El **remuestreo** es el proceso de hacer coincidir las **tasas de muestreo.**

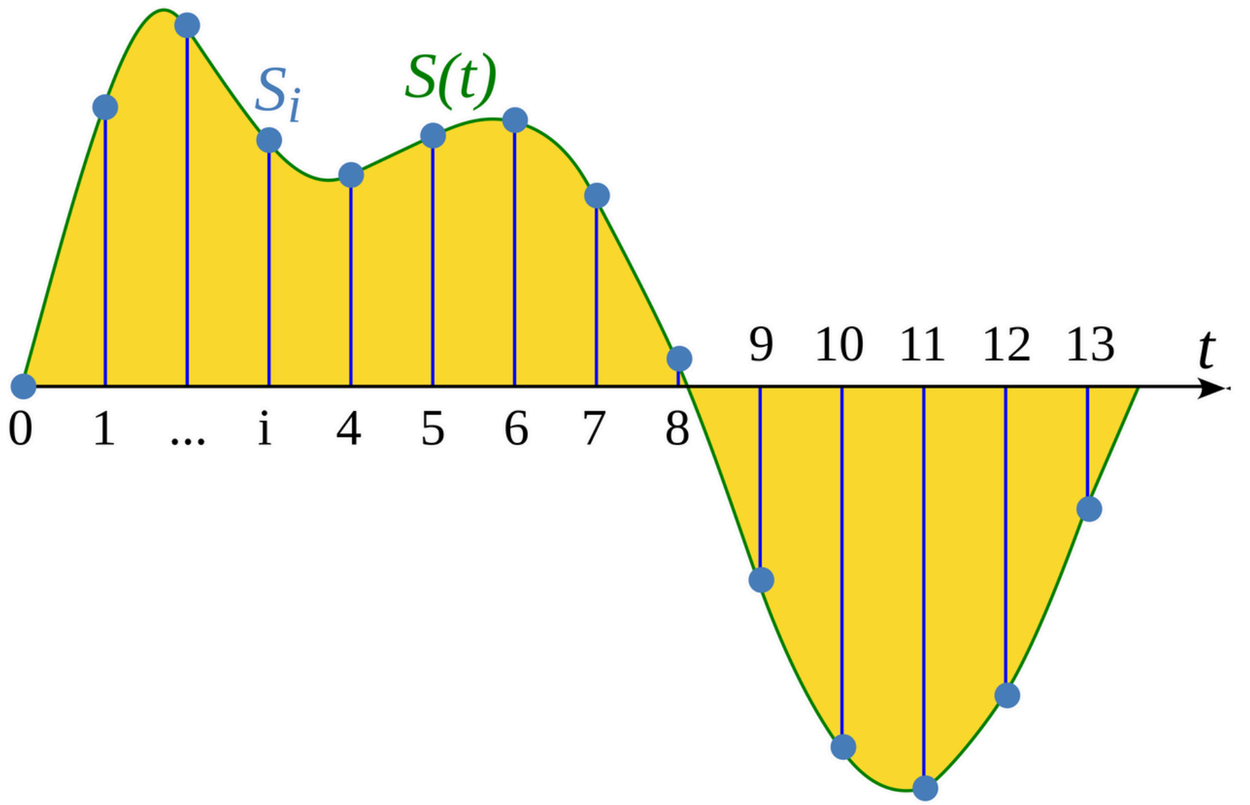


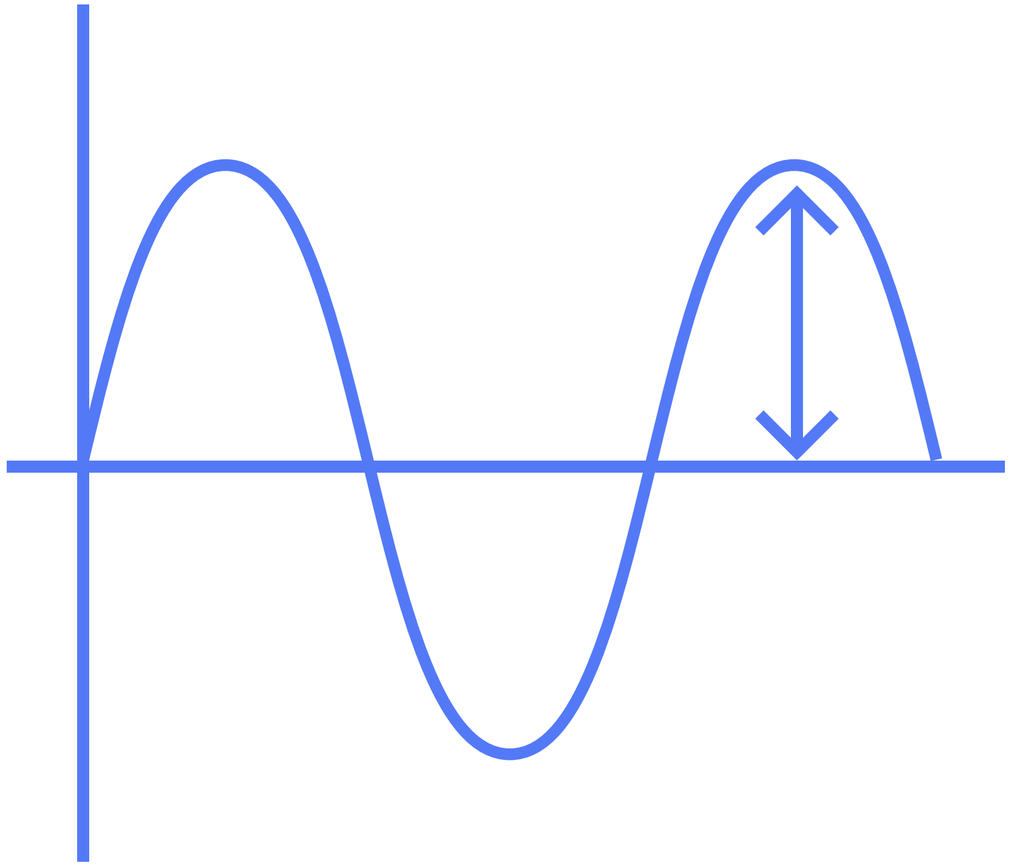
Imagen de wikipedia

AMPLITUD

El sonido se produce por cambios en la presión del aire a frecuencias audibles para el ser humano. La amplitud de un sonido describe el nivel de presión sonora en un instante dado y se mide en decibelios (dB).

PROFUNDIDAD DE BITS

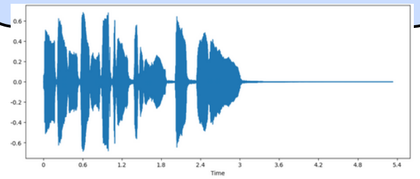
La profundidad de bits de la muestra determina con cuánta precisión puede describirse este valor de amplitud.



VISUALIZACION DEL AUDIO

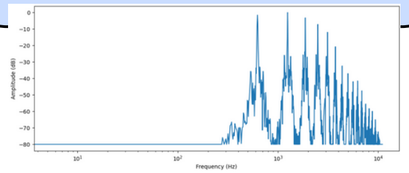
En forma de onda

Traza los valores de la muestra a lo largo del tiempo e ilustra los cambios en la amplitud del sonido.



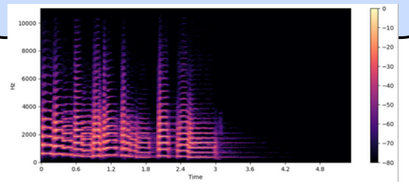
El espectro de frecuencias

El espectro se calcula mediante la transformada discreta de Fourier o DFT. Describe las frecuencias individuales que componen la señal y su intensidad.



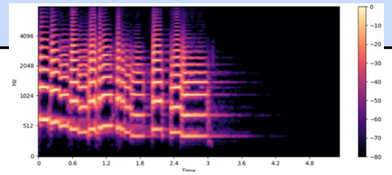
Espectograma

Un espectrograma representa el contenido de frecuencia de una señal de audio a medida que cambia con el tiempo.



Mel-Spectograma

El espectrograma Mel es un espectrograma en el que las frecuencias se convierten a la escala Mel.



PREPROCESSING

Remuestreo de los datos de audio

Filtrar el conjunto de datos

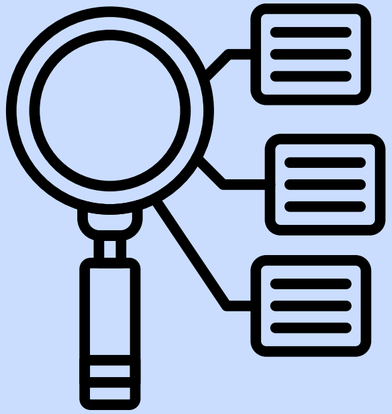
Conversión de los datos de audio a la entrada prevista del modelo

Datasets

Cargar un dataset, explorarlo y datasets en huggingface

APLICACIONES DE AUDIO USANDO IA

Clasificación de audio

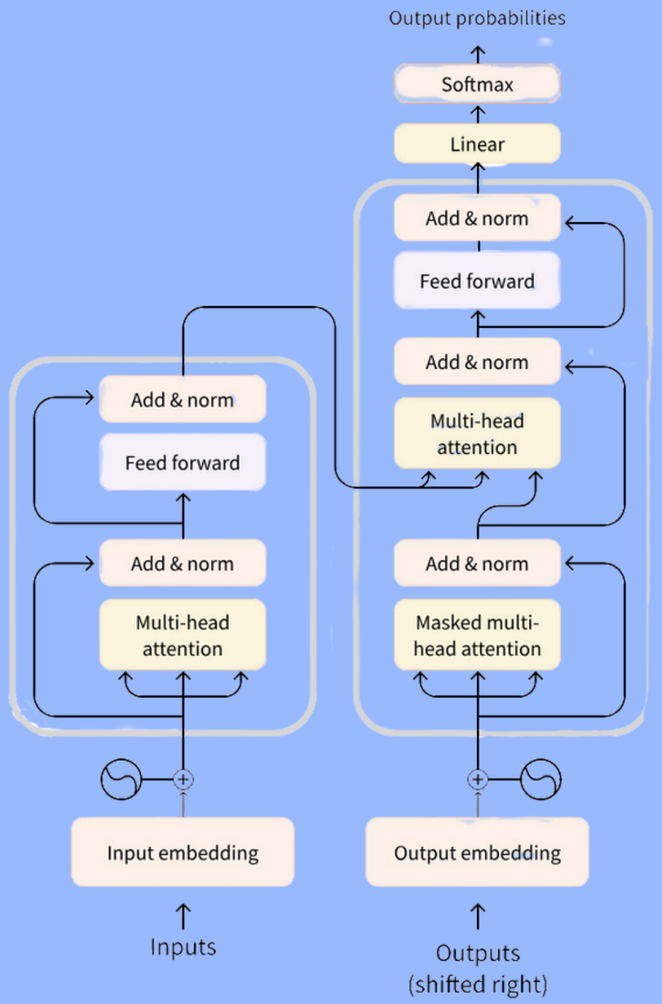


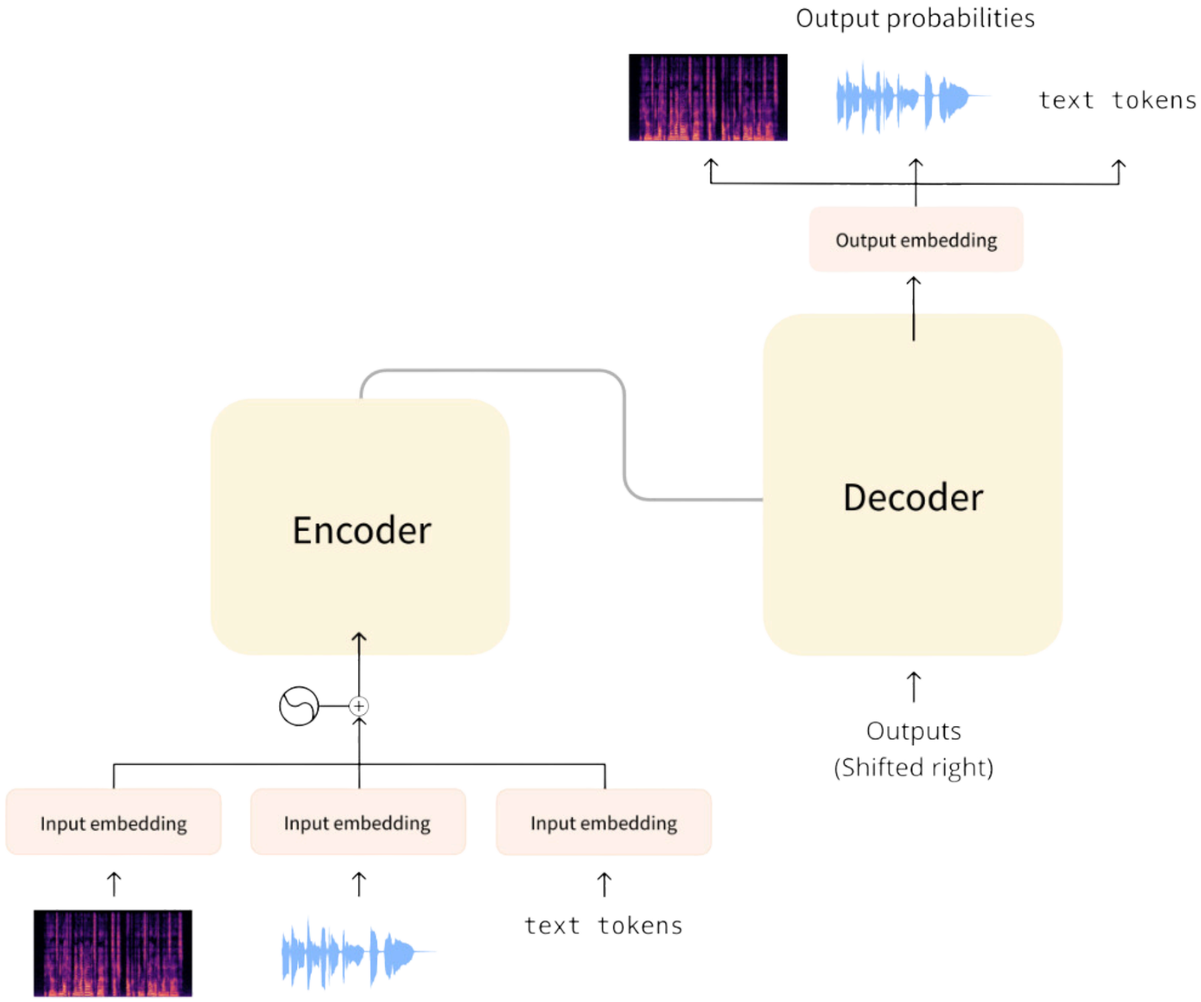
Reconocimiento automático de voz

Speaker diarization

Texto a voz

TRANSFORMERS



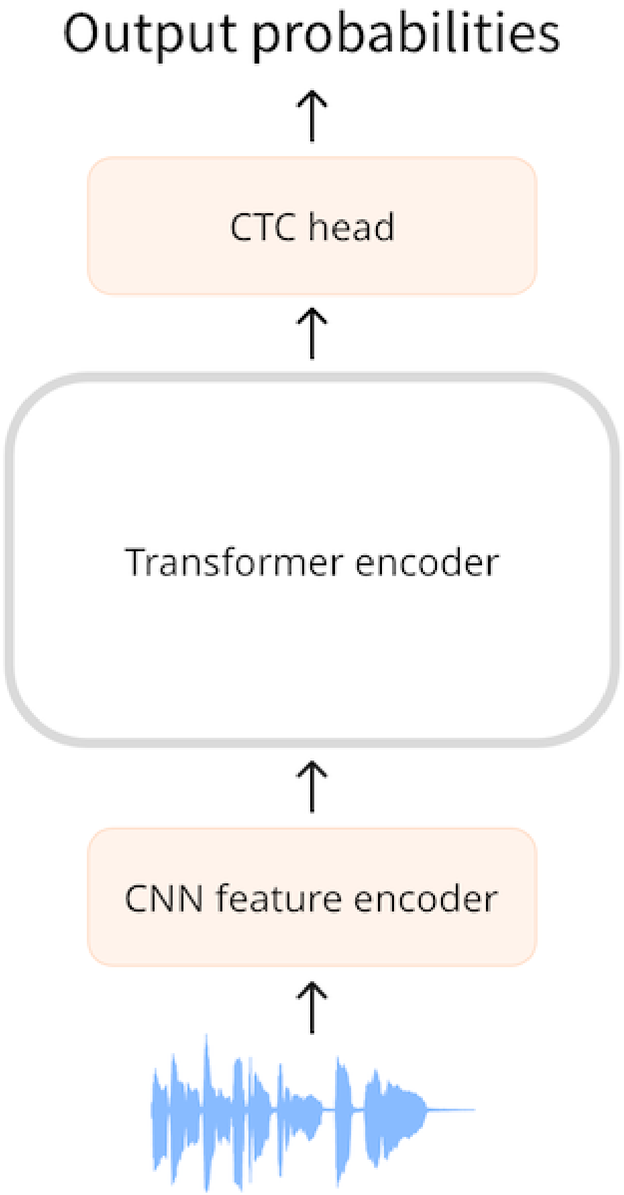


Reconocimiento automático del habla (ASR): La entrada es el habla, la salida es el texto.

ARQUITECTURA CTC (CONNECTIONIST TEMPORAL CLASSIFICATION)

El codificador lee la secuencia de entrada (en forma de onda) y la mapea en una secuencia de estados ocultos, también conocidos como **output embeddings.**

Con un modelo CTC, aplicamos un mapeo lineal adicional en la secuencia de estados ocultos para obtener predicciones de etiquetas de clase. Las etiquetas de clase son los caracteres del alfabeto (a, b, c, ...).



HAY UN PROBLEMA...

En el habla, no conocemos la alineación de las entradas de audio y las salidas de texto. Sabemos que el orden en el que se habla es el mismo que el orden en el que se transcribe el texto (la alineación es la llamada monotónica), pero no sabemos cómo se alinean los caracteres de la transcripción con el audio. Aquí es donde entra en juego el algoritmo CTC.

ALGORITMO CTC

Predecir un carácter cada x tiempo

Por ejemplo, en el modelo wav2vec2 predice un carácter cada 20 ms.

Uso de un token especial

Se usa un token especial que hace parte del vocabulario para delimitar grupos de caracteres.

Limpieza de la salida

Por último, se usa el token especial para remover caracteres repetidos y se elimina el token especial.

Paso adicional

Como adicional a la arquitectura, se le puede agregar un modelo NLP para corregir posibles errores de deletreo

EJEMPLO ALGORITMO CTC

Predecir un carácter cada x tiempo

ERRRRORR

Uso de un token especial

_ER_RRR_ORR

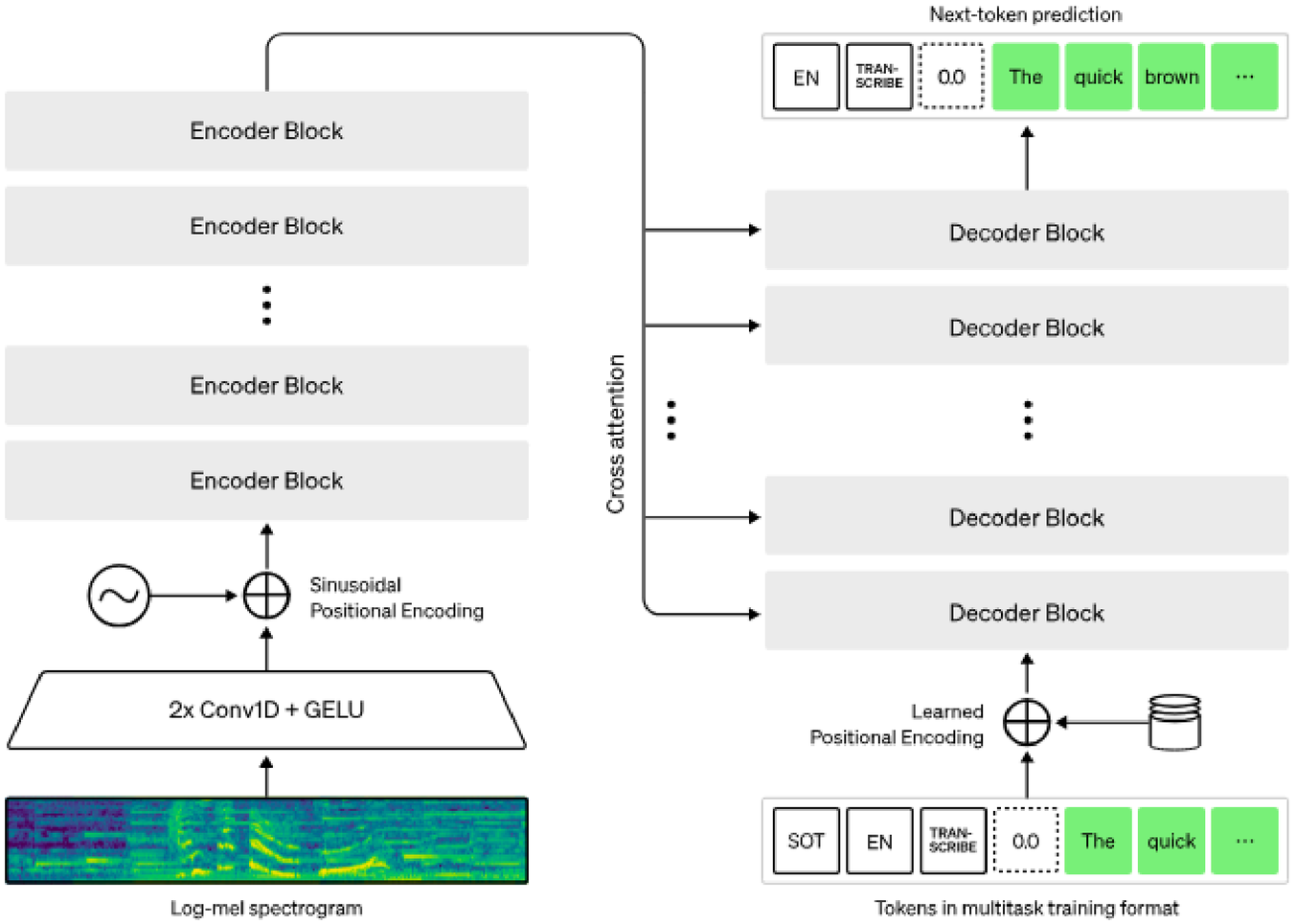
Limpieza de la salida

_ER_R_OR

ERROR

ARQUITECTURA SEQUENCE-TO-SEQUENCE

Whisper



METRICAS DE EVALUACIÓN

En el habla, no conocemos la alineación de las entradas de audio y las salidas de texto. Sabemos que el orden en el que se habla es el mismo que el orden en el que se transcribe el texto (la alineación es la llamada monotónica), pero no sabemos cómo se alinean los caracteres de la transcripción con el audio. Aquí es donde entra en juego el algoritmo CTC.

Sustitución (S)

Cuando transcribimos una palabra incorrecta en nuestra predicción

Inserción (I)

Cuando añadimos una palabra extra en nuestra predicción

Borrados (D)

Cuando eliminamos una palabra en nuestra predicción

WER (WORD ERROR RATE)

La métrica de la tasa de error de palabra (WER) es la métrica «por defecto» para el reconocimiento del habla. Calcula las sustituciones, inserciones y supresiones a nivel de palabra. Esto significa que los errores se anotan palabra por palabra.

$$WER = \frac{S + I + D}{N}$$

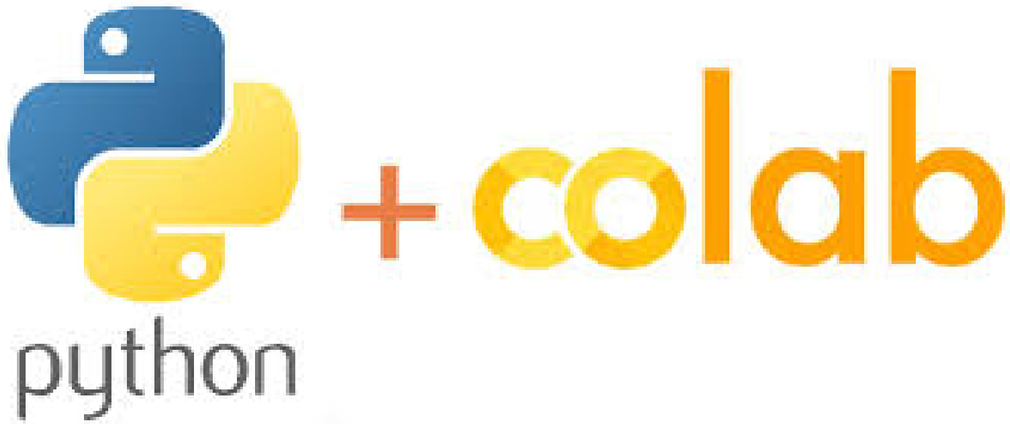
CER (CHARACTER ERROR RATE)

El índice de error de caracteres (CER) evalúa los sistemas a nivel de caracteres. Esto significa que dividimos las palabras en caracteres individuales y anotamos los errores carácter por carácter.

$$CER = \frac{S + I + D}{N}$$

EJEMPLO DE MODELO ASR USANDO FINE-TUNING

Fine-Tuning ASR model.ipynb



ACTIVIDAD:

Tomar el ejemplo que se explico y realizar un fine tunning usando el idioma español (consejo usen otro dataset)

THANK YOU FOR LISTENING!