Universidad del Valle
Maestría en Analítica e Inteligencia de Negocios
Analítica de Datos en Salud

TAREA No. 2: APLICACIÓN DE EXTRACCIÓN DE INFORMACIÓN

Para desarrollar esta actividad usted debe crear una herramienta de extracción de información que permita procesar automáticamente historias clínicas de cáncer de mama. Debe integrar un modelo de cancer de mama con un modelos de detección de la negación e incertidumbre. i

Entregable 1: Un script documentado en Python donde se carga el modelo **NER** de cancer de mama y se lo usa para extraer entidades. Usted debe procesar un conjunto de historias clínicas con el modelo NER. Las historias clínicas y el modelo NER ya entrenado se encuentran en los siguientes enlaces.

Historias Clínicas:
https://drive.google.com/drive/folders/14rvXajKquTwvs2O1ASDymy_uN7ZsbXEQ

Modelo NER en HugginFace:
`"anvorja/xml-roberta-large-finetuned-sp-ner-mama-biomedical-corregido"`

Entregable 2: Un script documentado en Python donde se muestre el proceso de validación de la negación e incertidumbre. Este script debe permitir cargar el modelo previamente entrenado y usarlo para detectar aquellas entidades que están negadas, afirmativas o con incertidumbre.

Modelo Negación/Intertidumbre HugginFace:

https://huggingface.co/JuanSolarte99/bert-base-uncased-finetuned-ner-negation_detection_NUBES

Entregable 3:

Integrar los modelos anteriores (NER, Negación) en un sript que permita procesar las historias clínicas y producir una base de datos estructurada. La base de datos es un archivo CSV que contiene las siguientes columnas:

patient_id: Número registro del paciente. Es una identificación
sentence: La oración donde se encontró la entidad.
NER: La entidad extraída dentro de la oración
Estado: Puede tomar 3 valores, Afirmativa, Negada, Sospechosa