Módulo 8

Sesión N° 2

ACTIVIDAD:

Detectando similitud y términos clave en textos clínicos breves

- Objetivo: Aplicar un pipeline de procesamiento de texto con SpaCy, NLTK y TfindfVectorizer, evaluando su impacto en la calidad de los datos y preparación para tareas de clasificación o agrupación.

Instrucciones:

1. Simula o descarga un dataset pequeño de notas clínicas (mínimo 10 entradas).
2. Aplica las siguientes transformaciones en orden:
   - Limpieza básica: minúsculas, remover signos, números, correos y URLs.
   - Tokenización y lematización con spaCy.
   - Eliminación de stopwords (usa las de spaCy o NLTK).
   - Vectorización con TfidfVectorizer.
3. Visualiza los términos más relevantes por documento.
4. Compara el corpus original vs preprocesado en términos de longitud media, vocabulario y repetición de palabras.

Indicaciones

- Tipo de entregable: Notebook en Google Colab con explicación paso a paso.
- Tiempo estimado: 2 horas.
- Cantidad de estudiantes: Individual.
- Evaluación: Calidad del preprocesamiento, visualización de términos, explicación clara y código comentado.

Anexo – Fragmento de texto de ejemplo

Paciente 001: Consulta por cefalea persistente y mareo leve. No refiere fiebre.
Paciente 002: Se presenta con cuadro de diarrea aguda y malestar general.
...