Trabajo Práctico Nº 2

Tema: Introducción – 2da parte

Fecha Inicio: 07/04/2026 **Fecha de Entrega:** 21/04/2026

Actividades:

El texto a continuación pertenece a la introducción de un trabajo (**Texto1**):

> *El Machine Learning es un campo de la inteligencia artificial que está impactando últimamente en todas las áreas del conocimiento. Las áreas de las ciencias sociales, en especial la educación, no es ajena a ella, por tanto, se realiza una revisión sistemática de la literatura sobre aquellas técnicas y aplicaciones del Machine Learning e inteligencia artificial en Educación. La falta de conocimientos y habilidades de los educadores en Machine Learning e inteligencia artificial limita la implementación óptima de estas tecnologías en la educación. El objetivo de este trabajo es identificar las oportunidades de mejora de los procesos de enseñanza-aprendizaje y la gestión educativa en todos los niveles del contexto educativo a través de la aplicación de Machine Learning e inteligencia artificial. Las bases de datos utilizadas para la búsqueda bibliográfica fueron Web of Science y Scopus, la metodología aplicada se basó en la declaración PRISMA para la obtención y análisis de 55 artículos publicados en revistas de alto impacto entre los años 2021 y 2023. Los resultados mostraron que los estudios trataron un total de 33 técnicas de Machine Learning e inteligencia artificial y múltiples aplicaciones que fueron implementadas en contextos educativos en niveles de educación primaria, secundaria y superior en 38 países. Las conclusiones mostraron el fuerte impacto que tiene el uso de Machine Learning e inteligencia artificial. Este impacto se ve reflejado en el uso de diferentes técnicas inteligentes en contextos educativos y el aumento de investigaciones en escuelas de secundaria sobre inteligencia artificial.*

1) Empleando la librería NLTK de Python, elimine las *stop_words* empleando el idioma español, *tokenize* el texto anterior, y muestre el resultado con la frecuencia de cada término, ordenado por frecuencia descendente (además del listado, muestre un gráfico con los 20 términos/tokens más frecuentes).

2) Del texto a continuación, aplique el proceso de eliminación de *stop_words* en inglés y *tokenización*, a continuación emplee el proceso de *Stemming* con los algoritmos de *Porter* y *Lancaster*, comparando los resultados de los dos procesos encolumnados(**Texto 2**):

   Information retrieval is the process of obtaining relevant information from a collection of data. It involves searching for and retrieving information from various sources, such as databases, the Internet, and digital libraries. Information retrieval is a vital aspect of many fields, including business, education, and healthcare. In recent years, technological advances have led to the development of sophisticated information retrieval systems that use artificial intelligence and machine learning algorithms to provide more efficient and accurate results. These systems can understand natural language queries and retrieve information from large and complex data sets. As the amount of data available continues to grow exponentially, the need for effective information retrieval systems becomes increasingly important. Organizations are constantly seeking ways to improve their information retrieval processes to gain a competitive edge and make better-informed decisions. With the right tools and

strategies, information retrieval can provide valuable insights and help drive success in various industries.

3) Aplique el proceso de *Stemming* para el **Texto 1** y muestre el resultado. Advierta si los algoritmos de *Porter* y *Lancaster* en NLTK poseen la implementación para el idioma español, sino es así, aplique otro algoritmo que si la posea.

4) Del primer párrafo del **Texto 1**, obtenga 2-gramas y 3-gramas de palabras, muestre los resultados en cada caso.

5) Empleando el corpus Brown de NLTK, detokenize el archivo ***cg58***.

   A. Tokenize en oraciones.
   B. Muestre las primeras 10.

6) Realice paso a paso el preprocesamiento del texto obtenido en el punto anterior, ello incluye:

   A. Eliminación de ruido
   B. Tokenización
   C. Normalización
   D. Eliminación de palabras vacías
   E. Obtener un listado de las 50 palabras más frecuentes
   F. Stemming. Obtener un listado de las 50 palabras más frecuentes
   G. Lematizacion. Obtener un listado de las 50 palabras más frecuentes
   H. Lematizacion indicando el PoS para los verbos.
   I. Realizar una representación tabular de los primeros 30 tokens indicando la palabra normal, realizado el stemming, lematización y lematización con PoS (verbos)