Trabajo Práctico Nº 2

Tema: Introducción – 2da parte

Fecha Inicio: 08/04/2025 **Fecha de Entrega:** 22/04/2025

Actividades:

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

> *Los sistemas recomendadores son herramientas enfocadas a ayudar a los usuarios a obtener aquella información que mejor se corresponda con sus intereses y preferencias. Mientras que un buscador habitual se centra en encontrar aquello que el usuario solicita, un sistema recomendador ayuda al usuario a tomar una decisión, que puede ser la compra de un producto en un portal de comercio electrónico, la lectura de un libro, la revisión de un artículo científico, el acceso a una página web en específico, o el estudio de determinado recurso educativo en una plataforma virtual de aprendizaje.*
>
> *La clasificación más popular de los sistemas recomendadores está asociada al algoritmo que emplean para realizar la tarea de minería correspondiente y divide a los métodos de recomendación en métodos de filtrado basado en el contenido, métodos de filtrado colaborativo, métodos de filtrado demográfico y métodos híbridos [1-4]. En adición, la literatura ha desarrollado tanto sistemas recomendadores para la sugerencia de ítems para usuarios individuales, como enfocados en grupos de usuarios []. Así, los sistemas recomendadores enfocados en grupos de usuarios se centran en la sugerencia de determinados tipos de ítems que tienden a ser consumidos en grupos y no por usuarios individuales, tales como programas de televisión y paquetes turísticos [].*
>
> *De manera general, los dominios iniciales de aplicación de los sistemas recomendadores han sido el e-commerce[]y el e-learning[,], aunque en los últimos tiempos estos sistemas están siendo aplicados a escenarios cada vez más diversos []. Así, son relevantes las aplicaciones de los sistemas recomendadores en escenarios de e-health[] y de e-tourism[], como dos contextos relevantes de particular importancia.*
>
> *Específicamente, resulta importante en los últimos años el desarrollo de sistemas recomendadores en el dominio del turismo[]. En este dominio existe mucha información en línea disponible y por tanto los sistemas recomendadores juegan un papel muy importante con vistas a ayudar a los usuarios en la toma de decisiones sobre qué paquete turístico comprar, qué instalación hotelera visitar, o qué recorrido turístico elegir, entre otras decisiones similares a tomar con vistas a lograr la satisfacción final del cliente [].*

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 ***cg73***.

   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)