ACTIVIDAD 2:
PRÁCTICA DE CLASIFICACIÓN DE TEXTOS.

Nota técnica CUNEF.
Esta nota técnica ha sido preparada por **Irene Cid Vega**, para ser utilizada como material de análisis y estudio. De ninguna forma pretende ilustrar recomendaciones de actuación sobre las empresas, las situaciones, o las personas mencionadas en el documento. Las propuestas conceptuales, opiniones y análisis que aparecen en este documento son responsabilidad del autor(es) y, por lo tanto, no necesariamente coinciden con las de CUNEF.

CUNEF – c/ Leonardo Prieto Castro, 2. 28040 Madrid. Tfno. (++34) 91 448 08 92. www.cunef.edu

Índice.

CUNEF

1. Objetivo

El objetivo de esta práctica es desarrollar un detector de mensajes spam, a partir de un histórico de mensajes etiquetados. El fichero es spam.csv, con las columnas separadas por comas. La etiqueta *spam* indica que el mensaje es spam, y la etiqueta *ham* que es un mensaje normal enviado por un usuario.

2. Guion de la actividad

1. Lea el contenido del fichero csv en un DataFrame.
2. Realice el pre-procesamiento que considere necesario. Puede utilizar funciones de la librería NLTK o spaCy, a su voluntad. Recomendamos una escritura modular del código, para poder hacer pruebas posteriormente, y contestar a las preguntas del punto 6.
3. Convierta el corpus de documentos en una matriz **TF-idf**.
4. Divida en un subconjunto de entrenamiento y otro de evaluación.
5. Llegados a este punto, realice modelos de entrenamiento al menos con algoritmos de clasificador bayesiano ingenuo, máquinas SVM y un modelo basado en árbol de decisión. Obtenga resultados de **accuracy** de la clasificación, así como las **matrices de confusión** para los tres modelos.
6. Conteste a las siguientes preguntas basándote en evidencias de código. ¿Tiene influencia en el resultado final el número máximo de features a utilizar? Prueba al menos dos números de features diferentes para los tres algoritmos y mide los resultados. ¿Modifica el resultado si no se eliminan las stop words? Pruébalo para los tres algoritmos y mide los resultados.
7. Imagínese que este entregable es una labor que le han solicitado en un entorno profesional, y que tiene que entregar esta documentación para comentar lo que ha descubierto (datos de entrada, rendimiento de los modelos, o cualquier descubrimiento que pueda ser importante). Comente los resultados obtenidos.

3. Formato y fecha de entrega

El entregable correspondiente a esta actividad será el notebook Python que ejecute las tareas anteriores.

La fecha límite de entrega para esta Actividad es el **19 de noviembre de 2023**