ACTIVIDAD 2:
PRÁCTICA DE CLASIFICACIÓN DE TEXTOS.

Última actualización: **02-Abril-2021**
CUNEF – c/ Leonardo Prieto Castro, 2. 28040 Madrid. Tfno. (++34) 91 448 08 92. www.cunef.edu

Índice.

1. Objetivo

El objetivo de esta práctica es realizar una clasificación de una serie de opiniones sobre un hotel, que están recogidas en el fichero **hotel.csv**. Este fichero contiene dos columnas. La primera, bajo el título **text**, contiene las opiniones a clasificar, mientras que la segunda, bajo el título **label**, contiene la puntuación otorgada. Las opiniones se agrupan según su calificación en un valor 5 y otro valor 3.

2. Guion de la actividad

1. Lea el contenido del fichero csv en un DataFrame. Se sugiere utilizar la función **pandas.read_csv**. Atención a la codificación de los datos entrantes.
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, viendo si se obtienen mejores resultados al utilizar stop-words, al realizar una extracción de formas canónicas, etc.
3. Divida el conjunto de documentos en un subconjunto de entrenamiento y otro de evaluación.
4. Convierta el corpus de documentos en una matriz **TF-idf**. Lo más cómodo es utilizar el vectorizador **TfidfVectorizer**, que forma parte de **sklearn**. ¿Tiene influencia en el resultado final el número máximo de features a utilizar?
5. Llegados a este punto, realice modelos de entrenamiento al menos con algoritmos de clasificador bayesiano ingenuo y máquinas SVM. Obtenga resultados de precisión de la clasificación, así como las matrices de confusión para ambos modelos.
6. Comente los resultados obtenidos. ¿Qué factores influyen? ¿Los resultados obtenidos son los esperados inicialmente? ¿A qué se deben estos resultados? Piense en la calidad del conjunto de datos con los que está trabajando

1.

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 **9 de Mayo de 2021**