**Procesamiento Natural del Lenguaje (NPL)**

Actividad 2

**Contexto**

La gerencia general ha aprobado el proyecto de la célula Data Science para el análisis de sentimientos, por lo que ahora tiene que iniciar el preprocesamiento de texto. Todo el equipo está muy motivado en poner manos a la obra, ya que es la oportunidad de poner en práctica lo aprendido. Las primeras tareas que se van a realizar son: tokenización, limpieza y reducción del vocabulario.

**Consignas**

1. Importación de librerías. Se realizará una segunda parte del análisis, a la cual vamos a llamar “Preprocesamiento de datos”. Primero vamos a importar todas las librerías que vamos a usar en el análisis, las que consideremos necesarias.
2. Limpieza, tokenización y reducción del vocabulario.
3. Se definirá una función similar a la vista en los ejemplos, pero debemos personalizarla, en otras palabras, no copiar los ejemplos textualmente, solo basarnos en ellos. En la misma función, podemos realizar la tokenización y limpieza, o se pueden crear dos funciones por separado; lo importante es que especifiquemos dónde aplicamos la tokenización y la realicemos con RegexpTokenizer. Para ello, se debe usar como parámetro el patrón de expresión regular que se diseñó en la consigna 4 de la actividad anterior.
4. Se codificará una función para lematización y otra para *stemming*, que lea el corpus de los comentarios, y entregue el corpus reducido a lemas y raíces respectivamente. Debemos basarnos en los ejemplos, pero no debemos copiarlos textualmente.
5. *Split* de la muestra. Realizaremos la división de la muestra en entrenamiento y test, utilizando el método train\_test\_split de la librería sklearn. Se destinará el 20 % de muestra para el test y el 80 % para el entrenamiento.



1. Distribución del *target* según muestra, *train* y test. Realizaremos un diagrama de barras de los porcentajes del *target* (bueno, malo, o info) según muestra *train* y test. Las distribuciones pueden no salir tan similares, debido a que la muestra test tiene baja materialidad. Lo importante es que por lo menos se mantenga el orden.

**Formato de entrega**

****

Se debe entregar esta parte en formato *notebook* de Jupyter. Escriba el código de cada parte solicitada en una celda “code”, y mantenga las salidas dentro del mismo documento. Sus comentarios y respuestas textuales las puede escribir en una celda tipo “markdown”.

**Cómo cargar la actividad**

**¡Felicitaciones por llegar al final de esta actividad!**

Para concluir tu entrega sigue estos pasos:

1. Haz clic en el botón "Enviar tareas" para comenzar el proceso de carga de tu archivo con las respuestas.

2. Adjunta el archivo que contiene tus respuestas a través de la interfaz proporcionada.

3. Una vez que hayas adjuntado el archivo, confirma tu entrega haciendo clic en el botón correspondiente.

Recuerda completar todos los puntos y tener cuenta las condiciones de entrega.

**¡Éxitos!**