Práctica 1D: Información mutua

Guillermo Ruiz

1. Preprocesa el texto y conviértelo a minúsculas, quita acentos y los siguientes caracteres:

   `";:,.\\-\"'/()[]¿?¡!{}~<>|«»--’"`

   Reemplaza las tabulaciones y saltos de línea con un espacio.

   Elimina las stopwords.

2. Encuentra las asociaciones de palabras más significativas usando la medida de información mutua para cada conjunto de datos. Cada conjunto de datos se procesan por separado.

Los archivos son:

- archivo_emojis_Proceso.csv
- archivo_emojis_Elpais.csv
- archivo_emojis_Elfinanciero.csv

3. Usa el contenido de la columna `title`. Los puedes leer con:

```
data = pd.read_csv("<ruta>/archivo_emojis_Elfinanciero.csv")
text = data['title'].to_list()
text = " ".join(text)
```

4. Crea un notebook que calcule la información mutua de un texto. Usa una ventana de dos palabras para calcular las probabilidades conjuntas $p(x, y)$.

5. Incluye las asociaciones más importantes respecto a la medida de información mutua para cada conjunto de datos proporcionado. Es decir, se deben de calcular todas las asociaciones entre el vocabulario único. Muestra las 10 asociaciones más importantes que aparezcan al menos 10 veces en el texto e indica su índice de información mutua.