Práctica 1

Procesamiento del lenguaje natural

Preprocesar el texto de una página web en Wikipedia

Enunciado

Se pide realizar tareas de preprocesamiento de texto a la entrada de la Wikipedia (tanto en castellano como más tarde en inglés) dedicada a **Alan Turing**.

El preprocesamiento del texto incluirá las siguientes tareas:

1. **obtención del contenido de la web,** en concreto del primer párrafo de la versión en castellano:

   > *"Es considerado uno de los padres de la ciencia de la computación y precursor de la informática moderna. Proporcionó una influyente formalización de los conceptos de algoritmo y computación: la máquina de Turing. Formuló su propia versión que hoy es ampliamente aceptada como la tesis de Church-Turing (1936).*

   Para extraer esta parte concreta del contenido de la página dedicada a Turing, puedes emplear la librería *BeatifulSoup* de Python, útil para hacer tareas de "scraping".

   Para introducirte en el uso de *BeatifulSoup*, dedica unos minutos a revisar el tutorial que tienes en esta página. No hace falta realizar los ejemplos que vienen en la misma, sino que basta con tenerla como manual de referencia para extraer el contenido indicado (sabiendo que lo que interesa obtener en este caso es un párrafo).

2. **aplicación de diferentes técnicas de preprocesamiento de texto**

   Una vez aislado el texto, cabe recorrer el "pipeline" apuntado en clase, pero de momento sólo en lo relativo a su primera fase, esto es, la de preprocesamiento del texto. En concreto, realiza las siguientes tareas:

a. Elimina todos los signos de puntuación y todo aquello que no sean palabras[1].
b. Cambia todas las palabras a minúsculas.
c. Tokeniza la cadena obtenida
d. Lematiza la cadena obtenida
e. Elimina todas las palabras irrelevantes

3. **Categorización del texto:**

Una vez preprocesado el texto, deseamos clasificarlo con arreglo a las siguientes categorías: narrativo, informativo, explicativo y descriptivo. La adscripción a una categoría u otra se hará a partir de un análisis morfológico de las palabras del texto (es decir, si estas son sustantivos, adjetivos, verbos o adverbios) y de la definición de unos umbrales.

Ejemplo: si el texto contiene 'x' sustantivos y el valor de 'x' supera cierto umbral (por ejemplo, 5), el texto se considerará que el texto es narrativo.

Los umbrales serán los siguientes:

- 5 verbos para textos narrativos
- 5 sustantivos para textos informativos
- 2 adjetivos para textos descriptivos
- 2 adverbios para textos explicativos

Si algún texto no se ajusta a ninguna categoría, se le asignaría como categoría "desconocida".

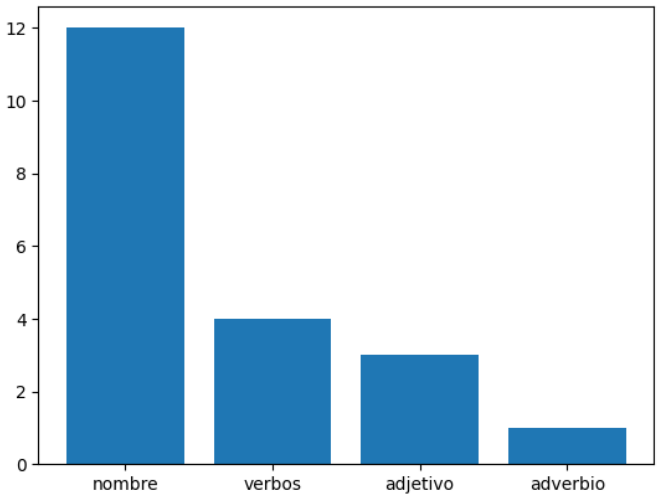
4. **Representación gráfica del análisis**:

Adapta el siguiente código de ejemplo para representar gráficamente el resultado:

```
import matplotlib.pyplot as plt
import numpy as np

x = np.arange(3)
plt.bar(x, height=[1,2,3])
plt.xticks(x, ['a','b','c']);
```

[1] Como aún no tienes el conocimiento necesario para definir patrones de reconocimiento de texto, puedes utilizar el siguiente → "\.|\(\d+\)|\:"



Guarda la práctica en un cuaderno Google Colab o Jupyter con el nombre "**P1_preprocesamiento_texto.ipynb**"