TEMA 2 – HOJA DE EJERCICIOS I

Se proponen diferentes ejercicios breves de procesamiento básico. **Todos los ejercicios propuestos en el listado se deben resolver utilizando la librería NLTK.**

Se proporcionan diferentes textos para trabajar en los ejercicios, aunque se pueden utilizar otros textos como entrada para realizar las comprobaciones que se consideren necesarias además de lo que se pide por defecto en cada uno de los ejercicios.

Ejercicios relacionados con análisis léxico/morfológico

1. Se pide implementar un código en Python para que sea capaz de tokenizar un texto en inglés.

   Punkt. Es el tokenizador de NLTK, el algoritmo que lleva a cabo la tokenización del texto que se le pasa por parámetro. Lo divide en sentencias por defecto. Un ejemplo de uso se ve a continuación:

```python
import nltk
from nltk.tokenize import PunktSentenceTokenizer

# Download the Punkt package
nltk.download('punkt')

# Sample text
text = "We met Miss. Tanaya Das and Mr.Rohan Singh today. They are pursuing a B.tech degree in Data Science."

# Initialize the PunktSentenceTokenizer
tokenizer = PunktSentenceTokenizer()

# Tokenize the text into sentences
sentences = tokenizer.tokenize(text)

print(sentences)
```

```
['We met Miss.', 'Tanaya Das and Mr.Rohan Singh today.', 'They are pursuing a B.tech degree in Data Science.']
```

   Dado el siguiente texto:

   The US Open will become a 15-day tournament in 2025, beginning on a weekend for the first time in the Open era.
   This year's main draw at Flushing Meadows will start on Sunday, 24 August and end on Sunday, 7 September.
   It becomes the latest Grand Slam to announce a Sunday start. The Australian Open expanded to a 15-day tournament in 2024, after the French Open took that decision in 2006.
   That leaves Wimbledon as the only remaining Slam event to retain the traditional Monday start.

In making the change, the US Open said the move would allow "more fan access than ever to the main draw following three consecutive years of record-breaking attendance".
The tournament estimates the expansion will allow access for an additional 70,000 spectators.
Men's and women's singles first-round matches will be played across the opening three days in New York, from Sunday to Tuesday.
The Australian Open took the decision to become a 15-day event in an attempt to reduce the number of late-night finishes at Melbourne Park.

El formato del resultado podría ser el siguiente (cada frase del texto original se muestra tokenizada en una línea diferente):

```
['The', 'US', 'Open', 'will', 'become', 'a', '15-day', 'tournament', 'in', '2025', ',',
['This', 'year', "'s", 'main', 'draw', 'at', 'Flushing', 'Meadows', 'will', 'start', 'o
['It', 'becomes', 'the', 'latest', 'Grand', 'Slam', 'to', 'announce', 'a', 'Sunday', 's
['The', 'Australian', 'Open', 'expanded', 'to', 'a', '15-day', 'tournament', 'in', '202
['That', 'leaves', 'Wimbledon', 'as', 'the', 'only', 'remaining', 'Slam', 'event', 'to
['In', 'making', 'the', 'change', ',', 'the', 'US', 'Open', 'said', 'the', 'move', 'wou
['The', 'tournament', 'estimates', 'the', 'expansion', 'will', 'allow', 'access', 'for
['Men', "'s", 'and', 'women', "'s", 'singles', 'first-round', 'matches', 'will', 'be',
['The', 'Australian', 'Open', 'took', 'the', 'decision', 'to', 'become', 'a', '15-day',
```

2. Hacer lo mismo que el ejercicio anterior, pero para un texto en español. ¿Hay alguna diferencia a la hora de tokenizar?

3. ¿Qué ocurre si se tokeniza utilizando split en lugar de word_tokenize? Haz pruebas sobre diferentes textos y compara la salida.

4. Dado un fichero de texto con contenido en inglés (“Data_Science.txt”) se quieren eliminar todos los signos de puntuación que aparezcan en el mismo.

5. Dado un fichero de texto con contenido en inglés (“Data_Science.txt”) se quiere saber cuáles son las 5 palabras más frecuentes en el texto, antes y después de eliminar los signos de puntuación. Además, se quiere saber cuántas veces aparece la palabra “data” en el texto.

Se puede consultar la api de NLTK: https://www.nltk.org/api/nltk.probability.html

6. Dado un fichero de texto con contenido en inglés (“Data_Science.txt”) se quiere obtener su contenido, dividirlo en frases y por cada una de ellas eliminar las palabras vacías (stop words) que contengan.

7. Se pide lo mismo que en el ejercicio anterior, salvo que hay que hacerlo para un texto en español (“Ciencia_de_datos.txt”).

8. Dado un fichero de texto con contenido en inglés (“Data_Science.txt”) se quiere hacer uso de expresiones regulares para encontrar palabras que cumplan unas determinadas condiciones en el texto, por ejemplo:
    a. Buscar las palabras que comiencen por ‘d’
    b. Buscar las palabras que comiencen por ‘s’, terminen por ‘e’ y cuya longitud total sea de 7 caracteres.

9. Dada una frase en inglés, obtener la etiqueta del Part-of-Speech para cada token.

10. Dado un fichero de texto con contenido en inglés (“Data_Science.txt”) se pide obtener los diferentes tokens del texto y por cada uno de ellos realizar stemming. Observa la salida, ¿Todos los *stems* obtenidos están en el diccionario?

11. Se pide lo mismo que en el ejercicio anterior, salvo que hay que hacerlo para un texto en español (“Ciencia_de_datos.txt”).

12. Dado un fichero de texto con contenido en inglés (“Data_Science.txt”) se pide obtener los diferentes tokens del texto y por cada uno de ellos lematizar.

13. Se pide lo mismo que en el ejercicio anterior, salvo que hay que hacerlo para un texto en español (“Ciencia_de_datos.txt”).

14. El lematizador basado en WordNet de NLTK no soporta lematizar en español. Podemos tratar de hacer nuestro propio lematizador para español utilizando un diccionario que tiene diferentes formas de palabras con su lema asociado. A partir de este diccionario, contenido en el fichero lemmatization-es.txt, se pide construir un lematizador y obtener los lemas del contenido del fichero “Ciencia_de_datos”.