Práctica de identificación de idioma y pre-procesamiento

Objetivo

El objetivo de esta práctica es utilizar los conocimientos adquiridos durante la clase para generar un programa Python que utilice las principales librerías de NLP para detectar el idioma de un documento y realizar una segmentación de frases. Posteriormente realizará una tokenización del mismo texto y por últimos determinará las formas normales de los términos presentes

1. Para comenzar genere una cadena de texto en la que introduzca un texto con el tamaño necesario para realizar la práctica. Debe tener al menos diez frases. Puede utilizar cualquier mecanismo habitual en Python para ello, es decir puede asignar directamente el valor pero también podrá leer un fichero de texto plano del disco de su ordenador, o leer un documento de una URL
2. Detecte el idioma del documento e imprima dicha información por la terminal. Recomendamos utilizar **textacy.text_utils.detect_language()**
3. Divida el documento en frases e imprímalas utilizando el tokenizador adecuado, que habrá tenido que crear previamente
4. Divida el documento en tokens e imprímalos. Utilice **nltk.word_tokenize()**
5. Por último obtenga las formas normales (lemas) de los anteriores tokens
6. En cada uno de los casos anteriores, explique el resultado obtenido. ¿Cambia la calidad de los resultados si el documento no está en español sino en inglés?