

Práctico 2: Tokenización / Morfología / DME

Ejercicio 1

- (a) Escriba una expresión regular en Python que permita generar hashtags
- (b) Escriba una expresión regular en Python que sustituya los hashtags reconocidos por (a) y los sustituya por el string *HASHTAG*

Ejercicio 2

En las siguientes oraciones, diga:

- cuántas palabras hay
 - cuántos tokens hay
 - cuántos tipos hay
- (i) Rocha ni siquiera tiene un hotel cuatro estrellas, ni una rambla, ni un shopping, ni un cine, ni una mesa de casino, ni una piscina termal marina.
 - (ii) Con el incendio grande del año pasado salieron de los montes de Punta Rubia cientos de personas.
 - (iii) El estado del arte es una modalidad de la investigación documental que permite el estudio del conocimiento acumulado (escrito en textos) dentro de un área específica.

Ejercicio 3

Calcule la distancia de edición entre REDUCCIÓN y DURACIÓN (utilizando distancia de Levenshtein), y muestre las ediciones posibles con esa distancia.

Ejercicio 4

Aplicando el algoritmo que calcula la Distancia de Mínima Edición de Levenshtein, vea qué palabra es la que se pretendió escribir en un texto cuando el corrector ortográfico encontró AAREOO. Las candidatas sugeridas son MAREO y APAREO. Justifique.