SIW PL3. Similitud entre textos.

Hugo Fonseca Díaz (UO258318)

Requerimientos

Se necesita Python 3.8 y la biblioteca *nltk* para ejecutar el script *similarity.py*.

Detalles de implementación

Este script está compuesto por dos clases:

- **BagOfWords**: vector que representa el conjunto de palabras de un texto.
- **Coefficient(coef_function)**: modela un coeficiente, para calcularlo recibe una función de coeficiente de un determinado tipo en el constructor. Esto se realiza mediante el uso del patrón de diseño *estrategia*.

Además de estas dos clases, se definen también una serie de funciones importantes:

- ***string_to_bag_of_words(text)***: coge una string y la procesa siguiendo una serie de pasos. Primero elimina los símbolos de puntuación (exceptuando los apóstrofes). A continuación obtiene los tokens y realiza las siguientes operaciones: los convierte a minúsculas, los lematiza y por último ignora las palabras vacías.
- ***load_lines(filename)***: obtiene las consultas y los textos de los ficheros *txt* proporcionados.
- ***find_best_text(query, texts, coefficient)***: devuelve el número del mejor texto para una determinada consulta.
- Las funciones de coeficiente (*dice*, *jaccard*, *cosine* y *overlapping*).