

Práctica 3 de Sistemas de información para la web

Ángel García Menéndez

Universidad de Oviedo

Introducción

El script de python está preparado para realizar el procesamiento de los textos y las consultas proporcionadas empleando los coeficientes exigidos.

Bibliotecas

El script requiere que se tiene instalada la biblioteca *nlTK*. El resto de dependencias de la misma se bajarán automáticamente de ser necesario.

Estructura

Dentro del script hay dos clases importantes:

- *BagOfWords*: Representando el conjunto de palabras de un texto, habiéndose tratado el mismo de forma pertinente.
- *Coefficient*: clase estrategia cuyos hijos implementan los diferentes coeficientes.

Funcionamiento

Invocando la orden de línea de comandos *python practica3.py* aplicaremos el conjunto de consultas ‘cran-queries’ sobre los documentos de la colección ‘cran-1400’. Como output se nos mostrará la consulta en sí, y el número del documento que mejor la satisface de acuerdo a cada coeficiente.

Otras consideraciones

A la hora de convertir las cadenas de texto a vectores de términos se han realizado las siguientes operaciones:

- Eliminación de símbolos de puntuación
- Tokenización
- Conversión a minúsculas
- Lematización
- Eliminación de palabras vacías

Asimismo, el output del script se ofrece en formato markdown para poder ser redirigido a un archivo para una lectura más cómoda. El procesamiento de los textos y su conversión a *bags of words* se realiza una sola vez en el programa para mayor optimización.