Búsqueda de categorías

Introducción

Cuando estudiamos la arquitectura de las distintas páginas, nos fijamos en que muchas de ellas estructuran las noticias según una serie de categorías. Estas categorías son un buen punto de partida para comenzar a etiquetar parte de la información que estamos extrayendo, pero como las páginas están estructuradas de forma distinta nos encontramos que dada una categoría (Ej: *Deportes*), es referenciada con distitnas etiquetas (Ej: *deportes, Deporte, depo, etc*) necesitamos un trabajo previo de aprendizaje para determinar que conjunto de etiquetas hacen referencia a una misma categoría.

Búsqueda de sinónimos

Como hemos visto anteriormente en la mayoría de casos un conjunto de palabras muy similares entre sí hacen referencia a una misma categoría, o nos encontramos con que se juntan palabras que hacen referencia a distintas categorías pero que en una o varias páginas están unidas (Ej: *saludyconsumo*). Las diferencias normalmente están en el uso de mayúsculas/minúsculas o singular/plural.

El primer paso del proceso es obtener una lista con todas las etiquetas únicas, para ello repasamos todas las páginas buscando las etiquetas que usa cada una y aplicamos la distancia de Levenshtein.

Distancia de Levenshtein

La distancia de Levenshtein es una algoritmo que mide el número de pasos que hay que dar para llegar de una palabra A a una palabra B. Los posibles pasos son sustituir, insertar o eliminar una letra. Si aplicamos el algoritmo con las palabras general y generales obtenemos una distancia de dos, ya que tenemos que insertar dos letras para llegar de una palabra a otra.

Con este algoritmo recorremos la lista de categorías tomando una palabra y la comparamos con las siguientes, y si la distancia que indica el algoritmo es menor o igual a 3, consideramos que son sinónimos y que hacen referencia a la misma categoría. Las relaciones que vamos estableciendo las almacenamos en una base de datos.

Casos especiales

Con el proceso anterior podemos hacer grupos con la mayoría de las etiquetas pero aun quedan palabras que no podemos asociar a ningún grupo.

Una opción es el caso de encontrarnos con una categoría que en realidad podemos separar en dos etiquetas, para classificar esta etiqueta buscamos otras cadenas que estén contenidas en la etiqueta compuesta. En el caso de *saludyconsumo* tenemos que salud y consumo forman parte de la etiqueta por lo que podemos decir que *saludyconsumo* responde a estas dos categorías.

Otro caso especial son el de aquellas etiquetas que no se parecen, según el criterio establecido, a ninguna otra palabra de la lista, bien sea por que la etiqueta es un acrónimo, una abreviatura o una palabra totalmente distinta pero que representa el mismo concepto de una determinada categoría. Estos casos se han añadido a alguna categoría de manera manual.