Tarea del Tema 5: Estudio de diferentes representaciones vectoriales

A partir de un pequeño corpus de páginas web (puedes descargar un conjunto de documentos HTML a tu elección), el objetivo de esta tarea es crear las representaciones de los documentos de dicho corpus dentro del *modelo de espacio vectorial* y usando, para cada uno de ellos, al menos dos de las funciones de pesado presentadas en el tema, una de carácter local y otra global. Deberán entregarse también documentos de texto con la información necesaria para generar las representaciones (el vocabulario, el fichero invertido y las propias representaciones).

Para generar las representaciones deberás aplicar las siguientes fases:

- **Análisis léxico**; seleccionando los rasgos con los que generar las representaciones. Deberás considerar los espacios en blanco como separadores entre cadenas. El resto de decisiones (eliminación de signos de puntuación, caracteres raros, ...) se dejan a tu elección.
- **Eliminación de stop-words**. Deberás buscar una lista de stop-words correspondiente a la lengua en la que estén escritos los documentos que hayas seleccionado y aplicarla a la fase de eliminación de palabras vacías.
- **Truncado**. Aplicar las reglas del algoritmo de Porter para textos en inglés (http://tartarus.org/martin/PorterStemmer/). En caso de que prefieras emplear documentos escritos en otra lengua, deberás buscar otro algoritmo de *stemming* y aplicar sus reglas de truncado (http://snowball.tartarus.org/texts/stemmersoverview.html).
- **Aplicación de dos funciones de pesado**, una de carácter local y otra global.

Finalmente deberás entregar una pequeña memoria que contenga una sección para cada uno de los puntos previos, y donde se expliquen razonadamente, y en cada caso, las decisiones que has ido tomando.

Nota:

Tanto el Vocubulario como las diferentes representaciones deberán generarse en un documento TXT, donde el formato de representación podrá ser:

- Vocabulario: $\mathbf{term_1}$ $\mathbf{term_2}$ … $\mathbf{term_N}$ (en una sola línea y separados por espacios/tab)
- Representaciones: $\mathbf{pesoTerm_1}$ $\mathbf{pesoTerm_2}$ $\mathbf{pesoTerm_N}$ (en una sola línea y separados por espacios/tab)
- Fichero invertido: $\mathbf{term_1 \rightarrow doc_1{:}TF_{1,1}}$ $\mathbf{doc_2{:}TF_{1,2}}$ $\mathbf{doc_3{:}TF_{1,3}}$ ... $\mathbf{doc_{10}{:}TF_{1,10}}$

  $\mathbf{term_2 \rightarrow doc_1{:}TF_{2,1}}$ $\mathbf{doc_2{:}TF_{2,2}}$ $\mathbf{doc_3{:}TF_{2,3}}$ ... $\mathbf{doc_{10}{:}TF_{2,10}}$

  ...

La información por término en una línea y donde $\mathbf{TF_{2,1}}$ representa la frecuencia de $\mathbf{term_2}$ en $\mathbf{doc_1}$

En caso de usarse otro formato, deberá detallarse claramente en la memoria.