Autotaging , un autocatalogador de documentos en función del contenido

Aplicativo, que con técnicas de minería de textos y algoritmos de clasificación aplicados, permite catalogar documentos en función de su contenido, en tres modos de trabajo: manual, aprendizaje supervisado y aprendizaje no supervisado.

En el panorama actual en el que la cantidad de información es ingente, en forma de textos, fotografía, vídeos, etc., se hace necesario tener a mano cualquier tipo de herramienta que ayude a la catalogación de toda esta enorme cantidad de contenidos; que permita a los profesionales no perderse en largos y tediosos procesos y llevar a cabo su tarea con la máxima eficiencia posible .

La filosofía de **Autotaging** es sencilla: se parte de un conjunto de categorías, y dentro de las mismas, de un conjunto de tags asignados. Y un tag puede estar más de una categoría.

Según el modo de trabajo, las categorías son introducidas por los usuarios expertos (modo **Manual**), o son definidas de forma automática por el sistema.

- **Aprendizaje supervisado:** Existen ya documentos catalogados con anterioridad (histórico) y el sistema, en base al contenido de dicha catalogación, extrae de los tags más representativos de cada categoría (se realiza la clasificación a partir de un árbol de decisión). Se creará un modelo de clasificación que se evalúa contra el resto de documentos no categorizados.

- **Aprendizaje no supervisad**o: En este caso, no existe ningún documento categorizado y el sistema, por similitud de contenido entre los documentos, los clasifica en clusters de conocimiento (segmentación de la información), en base a redes neuronales autoasociativas (SOM), que además, como veremos a continuación, permiten crear un mapa "GIS" que representa dicho conocimiento.

Una vez que los documentos ya están segmentados, se sigue el mismo proceso que en el aprendizaje supervisado para la extracción de los tags representativos asignados a la categorías. En este caso, las categorías tendrán nombres genéricos (categoriaA, B, etc.), y tendrá que ser el usuario experto, "a posteriori", el que les de nombres en función del contenido.

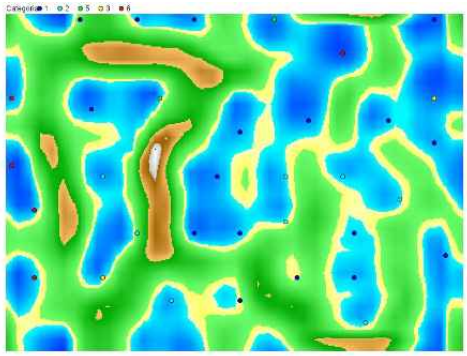
Vectorización de los documentos

La clasificación de los documentos se hace en función del contenido del mismo gracias a una técnica de minería de textos, denominada "vectorización",

Se seleccionan las palabras del documentos, se "lematizan" (se toma su raíz), y después se genera un vector que contiene tantas columnas como palabras tenga el conjunto de documentos y tantas filas como documentos.

El valor de cada columna es la frecuencia relativa inversa de la palabra con respecto al documento, es decir, lo relevante que es para el documento (cuantas más veces se repita, más relevante) e inversamente proporcional a lo repetida que sea en el resto de documentos (si en todos los documentos, en un entorno de contabilidad, por ejemplo, aparece "cuenta", dicho término no es relevante en general, para clasificar los documentos).

Por último, como los documentos están "vectorizados", por comparación de vectores y otras técnicas, se puede generar un mapa conceptual en donde los documentos más cercanos en un mapa serán los que estén en los mismos "valles", mientras que las montañas indicarán ausencia de documentos y separación entre segmentaciones "temáticas".



Los documentos pueden ser de cualquier tipo (word, excel, pdf, http, xml, etc.), y estar ubicados en directorios concretos, bases de datos, gestores de contenidos o en Web.