Tarea del Tema 7:

Weka. Data Mining with Open Source Machine Learning

Software in Java

En la siguiente página web http://www.cs.waikato.ac.nz/ml/weka/ el alumno podrá encontrar el Software Weka, una colección de algoritmos de aprendizaje automático desarrollados por la universidad de Waikato (Nueva Zelanda) e implementados en Java. Weka contiene las herramientas necesarias para realizar transformaciones sobre los datos, tareas de clasificación, regresión, clustering, asociación y visualización.

En esta tarea, el alumno deberá:

1) Descargar una colección de páginas web de la siguiente dirección de correo:
http://nlp.uned.es/~vfresno/banksearch/

La colección está formada por documentos HTML; aunque en apariencia sean archivos TXT, se trata de documentos HTML con una cabecera antes de la etiqueta inicial. En el nombre del archivo está codificada la categoría a la que pertenece (A,B,C,..., J)

2) Seleccionar aleatoriamente un 70% de los documentos de la colección anterior con los que se generará un vocabulario. A esta subcolección la denominaremos *traning set*, mientras que el 30% restante constituirá la subcolección *test set*.

3) Representar la colección completa empleando una función de pesado TF-IDF (o con cualquiera de las funciones de pesado utilizadas en la tarea 5) y con el vocabulario generado en el punto 2). Como la dimensión del vocabulario será muy elevada, deberá aplicarse un método de reducción; por ejemplo, eliminando aquellos términos que aparezcan en un número muy alto y muy bajo de documentos de la colección.

> **Nota 1**: El vocabulario deberá generarse a partir de los términos encontrados en el training set, y con este mismo vocabulario habrá que representar tanto el training como el test set. Esto significa que si un término aparece únicamente en páginas del test set, entonces no formará parte del Vocabulario.
> **Nota 2**: Dependiendo de los valores que se tomen como umbral, la dimensión será mayor o menor. Queda a la elección del alumno la decisión justificada de selección de estos umbrales.

4) Seleccionar y aplicar un algoritmo de clasificación de entre los que ofrece Weka, entrenando con el training set y clasificando el test set. La evaluación deberá realizarse usando las medida-F explicada en el tema.

Documentación a entregar:

Un informe en el que se describa brevemente el algoritmo seleccionado, se den los valores de clasificación obtenidos y se realice un análisis crítico de los resultados.