

Sistemas Inteligentes de acceso a la Información

Práctica – Construcción de categorizador basado en aprendizaje

1. Objetivo

El objetivo de la presente práctica es utilizar de manera aplicada los conocimientos adquiridos en teoría sobre la categorización automática de texto, con ayuda del paquete WEKA.

2. Desarrollo de la práctica

Se pide efectuar un experimento de categorización automática sobre datos del ODP usando el paquete WEKA. En particular, se ha de:

1. Confeccionar una colección de datos de entrenamiento y evaluación usando un conjunto de al menos 3 categorías y al menos 50 referencias del ODP en inglés.
2. Preparar los datos para su procesamiento con WEKA.
3. Aplicar los filtros adecuados para permitir el aprendizaje (incluyendo la conversión a archivo de índices y la selección de atributos).
4. Seleccionar y utilizar un rango de algoritmos de aprendizaje representativo (que incluya los vistos en clase que estén en WEKA).
5. Evaluar los resultados sobre la propia colección de entrenamiento.
6. Presentar y discutir los resultados obtenidos.

3. Requisitos de la práctica

Se debe entregar una memoria **breve** de la práctica, que incluya la descripción detallada del proceso, la elección motivada de los datos, filtros y algoritmos, y una discusión plausible de los resultados. La memoria será un archivo DOC o PDF. Los archivos se entregarán en un solo archivo comprimido ZIP terminado en OCX, llamado “grupoXXpractica02.zip.ocx”, adjunto a un mensaje de correo electrónico con la clave en el Asunto: “[SINAI] Grupo XX – Práctica 02”. La práctica se entregará a lo sumo el día 8 de enero de 2007.

La evaluación de la práctica tendrá en cuenta:

1. El cumplimiento de los requisitos arriba mencionados.
2. La efectividad obtenida.
3. La calidad de la discusión de los resultados.

Referencias

1. ODP – Open Directory Project – <http://dmoz.org>.
2. WEKA – Waikato Environment for Knowledge Analysis - <http://www.cs.waikato.ac.nz/ml/weka/>.