

Prácticas de Almacenamiento y Recuperación de la información

Construcción de un sistema de recuperación de información basado en el modelo vectorial

Curso 2004-2005

El alumno deberá construir un sistema de recuperación de información según el modelo vectorial en uno de los siguientes lenguajes: C, C++, Visual Basic, Pascal, Delphi o Perl (No se permite realizar el sistema en Java). El sistema debe poder correr en Windows o Linux (No es necesario que funcione en ambos).

Características

- El sistema debe estar basado en el modelo vectorial
- La función de semejanza será la función coseno entre los vectores pregunta y documento.
- Los archivos de la base de datos documental residirán en un directorio propio.
- El sistema debe aceptar nuevos documentos compuestos por un título, un identificador único y el texto del documento. El título del documento no es un identificador único del documento, y por tanto, varios documentos pueden tener el mismo título. El identificador de documento será único en el sistema.
- El sistema utilizará una lista de palabras vacías que se puede extraer del archivo stoplist.htm que se encuentra en la página web de la asignatura.
- El lenguaje de consulta del sistema consistirá en una lista de términos con pesos.
- El sistema debe facilitar la realimentación del usuario, permitiendo buscar documentos similares a uno dado.
- La respuesta del sistema se obtendrá por pantalla y en un archivo XML. Para la presentación en pantalla se mostrará una lista en la que aparezca el número de orden, título, identificador y relevancia. La relevancia es un valor en tanto por ciento.
- Para la salida XML hay que utilizar los archivos Resultados.dtd, Resultados.xml y Resultados.xsl que se encuentra en la página web de la asignatura.
- El sistema debe poder visualizar los documentos introducidos en la base de datos.
- La colección de documentos se encuentra en la página web de la asignatura.
- El trabajo se podrá realizar individualmente o por parejas.
- Se deberá entregar junto con el sistema, una pequeña memoria en la que se presente el diseño elegido, decisiones de diseño, características y estadísticas del funcionamiento del sistema. Se podrá añadir a esta documentación cualquier información que se considere oportuna.

Plazos de entrega

La documentación y el código correspondiente a esta práctica deberá ser entregada al profesor de prácticas, antes del viernes 21 de enero de 2005 a las 12:00h.

Los documentos necesarios para la práctica, se pueden encontrar en <http://alarcos.inf-cr.uclm.es/doc/ARI/ari.htm>